

37
1.9.70



Psychometrika

VOLUME XX—1955

JANUARY-DECEMBER

Editorial Council

Chairman:—HAROLD GULLIKSEN

Managing Editor:—

DOROTHY C. ADKINS

Editors:—M. W. RICHARDSON

PAUL HORST

Assistant Managing Editor:—

B. J. WINER

Editorial Board

R. L. ANDERSON

T. W. ANDERSON

J. B. CARROLL

H. S. CONRAD

L. J. CRONBACH

E. E. CURETON

ALLEN EDWARDS

MAX D. ENGELHART

WM. K. ESTES

HENRY E. GARRETT

BERT F. GREEN

J. P. GUILFORD

HAROLD GULLIKSEN

PAUL HORST

ALSTON S. HOUSEHOLDER

LYLE V. JONES

TRUMAN L. KELLEY

ALBERT K. KURTZ

FREDERIC M. LORD

IRVING LORGE

QUINN MCNEMAR

GEORGE A. MILLER

WM. G. MOLLENKOPF

FREDERICK MOSTELLER

GEORGE E. NICHOLSON

M. W. RICHARDSON

WM. STEPHENSON

R. L. THORNDIKE

LEDYARD TUCKER

D. F. VOTAW, JR.

S. S. WILKS

GODFREY THOMSON

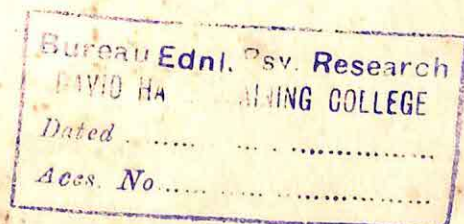
L. L. THURSTONE



PUBLISHED QUARTERLY

By THE PSYCHOMETRIC SOCIETY

AT 1407 SHERWOOD AVENUE
RICHMOND 5, VIRGINIA



Surge, Edm., by Research
DAVID L. STARRING COLLEGE
7.9.70
No. 8 - 373

Psychometrika

CONTENTS

SAMPLING FLUCTUATIONS RESULTING FROM THE SAMPLING OF TEST ITEMS	1
FREDERIC M. LORD	
SEPARATION OF DATA AS A PRINCIPLE IN FACTOR ANALYSIS	23
CHESTER W. HARRIS	
THE CHOICE OF AN ERROR TERM IN ANALYSIS OF VARI- ANCE DESIGNS	29
ARNOLD BINDER	
A RATIONAL CURVE RELATING LENGTH OF REST PERIOD AND LENGTH OF SUBSEQUENT WORK PERIOD FOR AN ERGOGRAPHIC EXPERIMENT	51
LEDYARD R. TUCKER	
A MEASURE OF INTERRELATIONSHIP FOR OVERLAPPING GROUPS	63
B. J. WINER	
AN EXTENSION OF ANDERSON'S SOLUTION FOR THE LA- TENT STRUCTURE EQUATIONS	69
W. A. GIBSON	
✓ A FACTOR ANALYSIS OF MENTAL ABILITIES AND PERSON- ALITY TRAITS	75
J. C. DENTON AND C. W. TAYLOR	
A TABULAR METHOD OF OBTAINING TETRACHORIC r WITH MEDIAN-CUT VARIABLES	83
GEORGE S. WELSH	
AN IBM METHOD FOR COMPUTING INTRASERIAL COR- RELATIONS	87
M. C. PAYNE, JR. AND L. STAUGAS	

VOLUME TWENTY

MARCH 1955

NUMBER 1



COOPERATIVE GRADUATE SUMMER SESSIONS IN STATISTICS

The University of Florida, North Carolina State College, Virginia Polytechnic Institute, and the Southern Regional Education Board are jointly sponsoring a series of cooperative summer sessions in statistics. The first session was held in 1954 at Virginia Polytechnic Institute. The second session will be held at the University of Florida from June 20 to July 29, 1955. A session is scheduled to be held at North Carolina State College in 1956, and another at Virginia Polytechnic Institute in 1957.

The sessions, based upon a recommendation of the Southern Regional Education Board's Advisory Commission on Statistics, will be of interest to (1) research and professional workers desiring instruction in basic statistical concepts, (2) teachers desiring training in modern statistics, (3) prospective candidates for graduate degrees in statistics, (4) graduate students in other fields wanting training in statistics, and (5) statisticians who wish to keep informed about advanced specialized theory and methods.

Each session lasts six weeks and each course carries approximately three semester hours of graduate credit. The program may be entered at any session; consecutive courses will follow in successive summers. The summer work may be applied as residence credit at any one of the cooperating institutions, as well as certain other institutions, in partial fulfillment of the requirements for a master's degree. The catalog requirements for the degree must be met at the degree-granting institutions. Doctoral candidates should consult with their institutions regarding the applicability of the courses. The faculty for the 1955 session will include:

R. L. Anderson, N. C. State College
D. B. Duncan, Univ. of Fla.
B. Harshbarger, Va. Poly. Inst.
C. E. Marshall, Okla. A. and M.
H. A. Meyer, Univ. of Fla.

G. E. Nicholson, Jr., Univ. of N. C.
P. J. Rulon, Harvard Univ.
W. L. Smith, Univ. of N. C.
D. B. South, Univ. of Fla.

Courses to be offered are:

Statistical Methods I
Statistical Methods II
 Design of Experiments
Statistical Theory I
Statistical Theory II
 Inference and Least Squares
Advanced Analysis I

Theory of Sampling
Theory of Statistical Inference
Statistical Research in
 Psychology and Education
Mathematics for Statistics
Seminar in Recent Advances
 in Statistics

The tuition fee is \$35 for the six-weeks term. The holder of a doctorate degree, upon acceptance, may register without the payment of tuition. Living and other expenses at the University are reasonable. Inquiries should be addressed to Professor Herbert A. Meyer, Statistical Laboratory, University of Florida, Gainesville, Florida.

Psychometrika

CONTENTS

- ✓ ESTIMATION AND TESTS OF SIGNIFICANCE IN FACTOR ANALYSIS 93
C. RADHAKRISHNA RAO
- RELIABILITY FORMULAS FOR NONCOMPLETED OR SPEEDED TESTS 113
LOUIS GUTTMAN
- A MATHEMATICAL MODEL FOR CONDITIONING 125
G. W. BOGUSLAVSKY
- TWO MODELS OF GROUP BEHAVIOR IN THE SOLUTION OF EUREKA-TYPE PROBLEMS 139
IRVING LORGE AND HERBERT SOLOMON
- ON THE DESIGN OF AUTOMATA AND THE INTERPRETATION OF CEREBRAL BEHAVIOR 149
STANLEY FRANKEL
- J. P. GUILFORD, *Psychometric Methods (2nd Ed.)* 163
A REVIEW BY BERT F. GREEN
- C. RADHAKRISHNA RAO, *Advanced Statistical Methods in Biometric Research* 165
A REVIEW BY HAROLD WEBSTER
- RAYMOND B. CATTELL. *Factor Analysis: An Introduction and Manual for the Psychologist and Social Scientist* 166
A REVIEW BY WILLIAM B. MICHAEL
- EDUCATIONAL TESTING SERVICE: *Kit of Selected Tests for Reference Aptitude and Achievement Factors* 169
A REVIEW BY H. J. EYSENCK

VOLUME TWENTY

JUNE 1955

NUMBER 2

Bureau Ednl. Psy. Research
DAVID H. A. KING COLLEGE
Dated
Accs No

NOMINEES FOR THE COUNCIL OF DIRECTORS
OF THE PSYCHOMETRIC SOCIETY

Two new members of the Council of Directors of the Psychometric Society are to be elected at the regular annual meeting of the Society in 1955. The following persons have been nominated:

OSCAR K. BUROS

ALSTON S. HOUSEHOLDER

JANE LOEVINGER

DAVID G. RYANS

Psychometrika

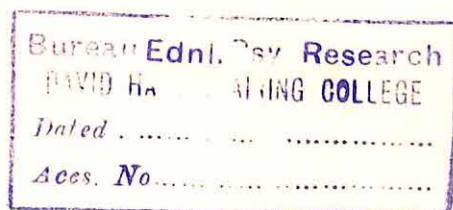
CONTENTS

SIR GODFREY THOMSON	171
L. L. THURSTONE	
A GENERALIZED SIMPLEX FOR FACTOR ANALYSIS . . .	173
L. GUTTMAN	
EQUATING TEST SCORES—A MAXIMUM LIKELIHOOD SOLUTION	193
F. M. LORD	
AXIOMS OF A THEORY OF DISCRIMINATION LEARNING .	201
F. RESTLE	
THE OBJECTIVE DEFINITION OF SIMPLE STRUCTURE IN LINEAR FACTOR ANALYSIS	209
L. R. TUCKER	
F-TEST BIAS FOR EXPERIMENTAL DESIGNS IN EDUCA- TIONAL RESEARCH	227
N. GOURLAY	
LEAST SQUARES ESTIMATES AND OPTIMAL CLASSIFICATION	249
H. E. BROGDEN	
AN IMPROVED METHOD FOR TETRACHORIC r	253
W. L. JENKINS	
BENJAMIN FRUCHTER, <i>Introduction to Factor Analysis</i>	259
A Review by C. WRIGLEY	
ANNE ANASTASI, <i>Psychological Testing</i>	261
A Review by W. M. DUROST	

VOLUME TWENTY

SEPTEMBER 1955

NUMBER 3





Psychometrika

CONTENTS

LOUIS LEON THURSTONE	263
J. P. GUILFORD	
PSYCHOMETRIC THEORY: GENERAL AND SPECIFIC	267
LEDYARD R. TUCKER	
F-TEST BIAS FOR EXPERIMENTAL DESIGNS OF THE LATIN SQUARE TYPE	273
NEIL GOURLAY	
CHARACTERISTICS OF TWO MEASURES OF PROFILE SIMILARITY	289
CHESTER W. HARRIS	
THE ESTIMATION OF THE DISCRIMINAL DISPERSION IN THE METHOD OF SUCCESSIVE INTERVALS	299
RAYMOND H. BURROS	
THE LAW OF COMPARATIVE JUDGMENT IN THE SUC- CESSIVE INTERVALS AND GRAPHIC RATING SCALE METHODS	307
H. J. A. RIMOLDI AND M. HORMAECHE	
A STATISTICAL MODEL FOR RELATIONAL ANALYSIS	319
R. DUNCAN LUCE, JOSIAH MACY, JR., AND RENATO TAGIURI	
REPORT OF THE TREASURER, PSYCHOMETRIC SOCIETY	329
REPORT OF THE TREASURER, PSYCHOMETRIC CORPORATION	329
INDEX FOR VOLUME 20	331
CUMULATIVE INDEX FOR VOLUMES 11 THROUGH 20	333
Author	334
Subject	338



SAMPLING FLUCTUATIONS RESULTING FROM THE SAMPLING OF TEST ITEMS*

FREDERIC M. LORD

EDUCATIONAL TESTING SERVICE

Sampling fluctuations resulting from the sampling of test items rather than of examinees are discussed. It is shown that the Kuder-Richardson reliability coefficients actually are measures of this type of sampling fluctuation. Formulas for certain standard errors are derived; in particular, a simple formula is given for the standard error of measurement of an individual examinee's score. A common misapplication of the Wilks-Votaw criterion for parallel tests is pointed out. It is shown that the Kuder-Richardson formula-21 reliability coefficient should be used instead of the formula-20 coefficient in certain common practical situations.

1. Introduction

Suppose that the same test is administered to a large number of separate groups of examinees, the groups being random samples all drawn from the same population; and suppose that some test statistic is computed separately for each sample of examinees. The value obtained for this test statistic will, of course, differ from sample to sample because of sampling fluctuations. The standard deviation of these values over a very large number of samples is the standard error of the test statistic *when examinees are sampled*. For convenience, this type of sampling will be referred to as *type-1* sampling.

On the other hand, suppose that a large number of forms of the same test are administered to the same group of examinees, each form consisting of a random sample of items drawn from a common population of items; and suppose that some test statistic is computed separately for each form of the test. Let us assume for theoretical purposes that the examinees do not change in any way during the course of testing, i.e., that there is no practice effect, no fatigue, etc. The value computed for the test statistic will still, of course, differ from form to form because of sampling fluctuations. The standard deviation of these values over a very large number of samples is the standard error of the test statistic *when the test items are sampled*. This type of sampling will be referred to as a *type-2* sampling. Test forms constructed by type-2 sampling will be called *randomly parallel forms* or *randomly parallel tests*.

Type-1 sampling fluctuations are familiar to everyone; type-1 standard

*Most of the work reported here was carried out under contract with the Office of Naval Research. The writer is indebted to Professor S. S. Wilks, who has checked over certain critical portions of a draft of this paper.

error formulas have long been available; they are sometimes incorrectly used in situations where sampling of test items is of crucial importance. Formulas for type-1 and type-2 standard errors may usually be readily distinguished on a superficial level by the following characteristics, which underscore the essential difference between them: type-1 standard errors are usually obviously proportional to some power (positive or negative) of the number of examinees in the sample and are usually much less obviously and simply related, if at all, to the number of items in the test; type-2 standard errors have the corresponding characteristic with respect to the number of items in the sample.

Section 2 of the present paper summarizes notation and lists type-2 standard error formulas without proof. Section 3 discusses two practical illustrative situations in which sampling of items is of crucial importance. Section 4 investigates the relation between the standard errors of individual examinees' scores and the Kuder-Richardson reliability coefficients, and reaches some important conclusions regarding the formula-21 coefficient. Section 5 discusses certain familiar formulas, including the Wilks-Votaw criterion for parallel tests, in relation to type-2 sampling formulas. Section 6 shows that the type-2 sampling distribution of most test statistics will be approximately normal when the number of test items is sufficiently large. Section 7 gives the derivation of the type-2 standard errors presented in section 2. Section 8, finally, discusses simultaneous sampling of items and examinees (type-12 sampling) and derives certain standard error formulas appropriate for this more complicated situation.

2. Notation and Summary of Formulas

In the present study, standard errors are obtained for the following test statistics:

t_a = the observed test score of examinee a , obtained by counting the number of items answered correctly on a single test.

\bar{t} = the mean of the scores obtained by the N examinees on a single test.

$$\bar{t} = \sum_a t_a / N.$$

s_t = the standard deviation of the scores obtained by the N examinees on a single test.

$$s_t^2 = \sum_a t_a^2 / N - \bar{t}^2.$$

r_{21} = the Kuder-Richardson reliability coefficient, formula 21.

$$r_{21} = \frac{n}{n-1} [1 - \bar{t}(n - \bar{t}) / ns_t^2].$$

r or r_{20} = the Kuder-Richardson reliability coefficient, formula 20.

$$r = \frac{n}{n-1} (1 - \sum_i s_i^2 / s_t^2)$$

(symbols explained in the succeeding list).

r_{ct} = the correlation of the test score with any external variable, c . $r_{ct} = s_{ct} / s_c s_t$.

Considerable care in defining notation must be taken here in order to avoid serious confusion. Additional symbols that will be used are listed below for easy reference.

x_{ia} = the "score" of examinee a on item i .

$$\begin{cases} x_{ia} = 1 & \text{if item answered correctly} \\ = 0 & \text{otherwise.} \end{cases}$$

n = the number of items in a single form of a test, i.e., in a single sample. The subscript i runs from 1 to n .

N = the number of examinees in a single group of examinees. The subscript a runs from 1 to N .

m = the number of items in a finite population of items.

p_i = the observed "difficulty" of item i for the N examinees tested.

$$p_i = \sum_a x_{ia}/N.$$

$q_i = 1 - p_i$.

z_a = the "proportion-correct score" of examinee a ; the proportion of the items in a single test answered correctly by examinee a . $z_a = t_a/n$.

\bar{z} , \bar{c} , etc. = the mean of the N values of z , c , etc.

$$\bar{z} = \sum_a z_a/N, \quad \text{etc.}$$

$M(p)$ = the mean of the n observed values of p_i for the n items in the test administered.

$$M(p) = \sum_i p_i/n.$$

s_c , s_z , etc. = the standard deviation of the N values of c , z , etc.

$$s_z^2 = \sum_a z_a^2/N - \bar{z}^2, \quad \text{etc.}$$

s_i = the standard deviation of x_{ia} for fixed i .

$$s_i^2 = \sum_a x_{ia}^2/N - (\sum_a x_{ia}/N)^2 = p_i q_i.$$

s_{ct} , etc. = the covariance (over examinees) of c and t , etc.

$$s_{ct} = s_c s_t r_{ct} = \sum_a (c_a - \bar{c})(t_a - \bar{t})/N.$$

s_{ic} , s_{iz} , s_{it} = the covariance (over examinees) of c_a , z_a , or t_a , respectively, with x_{ia} , for fixed i .

$$s_{it} = s_i s_t r_{it} = \sum_a (x_{ia} - p_i)(t_a - \bar{t})/N.$$

$s(p)$ = the standard deviation of the n observed values of p_i for the n items in the test administered.

$$s^2(p) = \sum_i p_i^2/n - M^2(p).$$

$s(s_{iz})$, $s(s_{it})$, etc. = the standard deviation of the n observed values of s_{iz} , s_{it} , etc. for the n items in the test administered.

$$s^2(s_{it}) = \sum_i s_{it}^2/n - (\sum_i s_{it}/n)^2.$$

$s(s_{ic}, s_{it})$ = the covariance (over items) of s_{ic} and s_{it} .

$$s(s_{ic}, s_{it}) = \sum_i s_{ic} s_{it}/n - (\sum_i s_{ic}/n)(\sum_i s_{it}/n).$$

r_{ic} , r_{it} , r_{iz} = the correlation of c_a , t_a , or z_a , respectively, with x_{ia} , for fixed i . $r_{it} = s_{it}/s_i s_t$.

TABLE 1
STANDARD ERRORS OF TEST STATISTICS

Statistic	Type 2 (Sampling test items)	Type 1 (Sampling examinees)
t_a	$\sqrt{\frac{t_a(n - t_a)}{n}}$	—
\bar{t}	$\sqrt{n} s(p)$	$\frac{s_t}{\sqrt{N}}$
s_t	$\frac{\sqrt{n} s(s_{it})}{s_t}$	$\frac{s_t}{\sqrt{2N}}$
r_{20}	$\frac{\sqrt{n}}{s_t^2} \sqrt{s^2(s_i^2) - 4(1 - r_{20})s(s_i^2, s_{it}) + 4(1 - r_{20})^2 s^2(s_{it})}$	*
r_{21}	$\frac{1}{\sqrt{n} s_t^2} \sqrt{(n - 2\bar{t})^2 s^2(p) + 4n^2(1 - r_{21})^2 s^2(s_{it}) - 4n(1 - r_{21})(n - 2\bar{t})s(p, s_{it})}$	*
r_{ct}	$\frac{\sqrt{n}}{s_t} \sqrt{\frac{1}{s_c^2} s^2(s_{ic}) - \frac{2r_{ct}}{s_c s_t} s(s_{ic}, s_{it}) + \frac{r_{ct}^2 s^2(s_{it})}{s_t^2}}$	$\frac{1 - r_{ct}^2}{\sqrt{N}}$

*Not known to writer.

It should be noted that all the statistics in the foregoing list are observed sample statistics relating to a given sample. There are two kinds of statistics listed, typified, in the simplest case, by $\bar{z} = \sum_a z_a/N$ and $M(p) = \sum_i p_i/n$. Population parameters have not been listed but will be designated, when needed, by the use of Greek letters. The following additional symbols, relating to the totality of all possible samples of test items (type-2 sampling), will be used.

- $E(x)$ = the *expected value* of x ; the arithmetic mean of the statistic x over all possible samples.
 $S.E.(x)$ = the standard error of the statistic x ; the standard deviation of the statistic x over all possible samples. $S.E.^2(x) = E(x^2) - [E(x)]^2$.
 $\text{var}(x)$ = the *sampling variance*. $\text{var}(x) = S.E.^2(x)$.
 $\text{cov}(x, y)$ = the *sampling covariance* of the statistics x and y over all possible samples.
 $\text{cov}(x, y) = E(xy) - E(x)E(y)$.

Table 1 summarizes the more important of the type-2 standard errors derived in the present paper. For purposes of comparison, the last column of the table, when appropriate, gives the corresponding usual type-1 formulas for the standard error for the case where the test scores are assumed to be normally distributed. The standard error formulas in both columns are large-sample formulas, in general, and observable sample statistics have been substituted for the corresponding population values throughout.

Type-12 standard errors are not listed here; their treatment is left for a special section.

3. Illustrative Examples and Discussion of the Standard Errors

Suppose that Form A of a certain 135-item test has been administered. Several parallel forms of this same test are to be administered in the future. Each form is administered to a different group of examinees. The groups of examinees may be considered as random samples drawn from the same population. Each group is so large that differences between groups due to sampling of examinees may be ignored. It is found that the mean, standard deviation, and Kuder-Richardson formula-20 reliability of the scores on Form A are 63.5, 21.5, and 0.95, respectively. How much may we expect the means to vary from form to form?

The required value of $s(p)$ could, of course, be determined directly from item analysis data. However, this value can be calculated, by means of (1), from the three numerical values given in the preceding paragraph. (1) is readily obtained by solving for $s^2(p)$ in Tucker's modification (9) of the usual formula for the Kuder-Richardson formula-20 reliability coefficient.

$$s^2(p) = \frac{s_t^2}{n} \left(\frac{n-1}{n} r_{20} - 1 \right) + \frac{l}{n} - \frac{l^2}{n^2}. \quad (1)$$

We find that $s^2(p) = .0538$.

The large-sample estimate of the type-2 standard error of the mean is found to be $S.E._2(\bar{t}) = 2.7$. (The subscript "2" is used here, and the subscript "1" is used below, to indicate type-2 and type-1 standard errors, respectively. Hereafter, type-2 sampling will be understood, unless otherwise specifically indicated.) If the same test were administered to random groups of 135 examinees, the type-1 standard error would be $S.E._1(\bar{t}) = 1.8$.

On the basis of the foregoing, we may expect that parallel forms of the test would not differ from each other in mean score by as much as $2\sqrt{2}S.E._2(\bar{t}) = 7.6$ points more than one time in twenty. If the parallel forms are carefully constructed by matching items from form to form on difficulty and item-test correlation rather than by random sampling of items, it may well be that the forms will not differ from each other as much as the foregoing formulas would indicate.

Suppose, for example, it is desired to investigate the relation of length of reading passage to validity in a reading comprehension test. The experimenter might well select at random from a pool of all available reading items of some specified difficulty level (a) a sample of all items based on passages containing more than 200 words and (b) a sample based on passages containing less than 100 words (it is assumed here that there is only one item per reading passage). He then places these items in random order and administers them to a group of examinees, obtaining separate scores for the long and for the short items. He computes the validity of each score, using some available criterion. If the two validity coefficients differ by little more than the type-2 standard error of their difference, it seems likely that the difference is attributable to chance fluctuations due to the sampling of items. If they differ by several times this standard error, the opposite conclusion may be reached; insofar as other uncontrolled experimental variables are ruled out, the difference may plausibly be attributed to length of reading passage.

A note of caution is necessary in using the type-2 standard error formulas. These formulas involve no assumptions beyond random sampling and large n ; however, *it is not at present known just how large an n is needed in any given case. The formulas in Table 1, therefore, should be used with some caution.* This is particularly true of the last three rows of the table, since the correlation coefficients given in the first column undoubtedly have sharply skewed distributions when n is small.

It should, finally, be noted that the assumption of random sampling of items cannot be expected to hold for speeded tests, and the formulas given in the present paper must be considered inapplicable.

4. Standard Errors of Measurement and Test Reliability

Table 1 gives a practical approximation to $S.E.(t_a)$ in terms of observed sample statistics; the rigorously accurate value, as shown in a later section, is

$$\text{S.E.}(t_a) = \sqrt{\frac{1}{n} \tau_a(n - \tau_a)}. \quad (2)$$

Here $\tau_a = E(t_a)$ is the true score of examinee a , i.e., the expected value of t_a over all randomly parallel forms of the test. [The expectation symbol, E , denotes the mean value over all type-2 samples; thus the operator E can be treated by the same rules as a summation sign.] The standard error of the score of an examinee is the standard deviation of the errors of measurement of his score (error of measurement = $t_a - \tau_a$). The average, taken over all examinees, of the squared values of such standard deviations of errors of measurement,

$$\frac{1}{N} \sum_a \text{S.E.}^2(t_a) = \frac{1}{N} \sum_a E(t_a - \tau_a)^2, \quad (3)$$

may appropriately be compared with the conventional "standard error of measurement" of test theory. This latter, which will be denoted by "S.E. Meas.," is likewise an average over all examinees. It is conventionally defined by the formula

$$\text{S.E. Meas.} = s_t \sqrt{1 - \text{reliability}}. \quad (4)$$

Specifically, it will now be shown that the squared standard error of measurement given by (3) is exactly equal to that which would be expected in (4) if the test reliability there were given by the Kuder-Richardson formula-21 coefficient (6). In our notation, this coefficient is

$$r_{21} = \frac{n}{n-1} \frac{s_t^2 - \bar{t}(1 - \bar{t}/n)}{s_t^2}. \quad (5)$$

The significance of the present proof is that it shows that *the Kuder-Richardson formula-21 coefficient (and, as will be seen, the formula-20 coefficient also) is no more nor less than a measure of the magnitude of type-2 sampling errors* (relative, of course, to the magnitude of true score differences).

Averaging (2) over all examinees, we find

$$\begin{aligned} \frac{1}{N} \sum_a \text{S.E.}^2(t_a) &= \frac{1}{nN} \sum_a \tau_a(n - \tau_a) \\ &= \frac{1}{N} \sum_a \tau_a - \frac{1}{nN} \sum_a \tau_a^2 \\ &= \bar{\tau} - \frac{1}{n} (\sigma_\tau^2 + \bar{\tau}^2). \end{aligned} \quad (6)$$

From (5) and (4), the expected value of the squared S.E. Meas. is

$$E[s_t^2(1 - r_{21})] = E\left[\frac{1}{n-1} (n\bar{t} - \bar{t}^2 - s_t^2)\right]. \quad (7)$$

In order to deal with (7) we first need expressions for $E(s_i^2)$ and $E(\bar{l})^2$:

$$\begin{aligned} E(s_i^2) &= E\left[\frac{1}{N} \sum_a (t_a - \bar{l})^2\right] \\ &= E\left[\frac{1}{N} \sum_a \{(t_a - \tau_a) + (\tau_a - \bar{\tau}) - (\bar{l} - \bar{\tau})\}^2\right]. \end{aligned} \quad (8)$$

After squaring and rearranging E and \sum signs,

$$\begin{aligned} E(s_i^2) &= \frac{1}{N} \left[\sum_a E\{(t_a - \tau_a)^2\} + E\left\{\sum_a (\tau_a - \bar{\tau})^2\right\} + NE\{(\bar{l} - \bar{\tau})^2\} \right. \\ &\quad + 2 \sum_a (\tau_a - \bar{\tau})E\{(t_a - \tau_a)\} - 2E\{(\bar{l} - \bar{\tau}) \sum_a (t_a - \tau_a)\} \\ &\quad \left. - 2E\{(\bar{l} - \bar{\tau}) \sum_a (\tau_a - \bar{\tau})\} \right]. \end{aligned} \quad (9)$$

Now the fourth and the last terms on the right vanish since $E(t_a - \tau_a)$ and $\sum_a (\tau_a - \bar{\tau})$ both equal zero. It is seen that we have, term for term,

$$E(s_i^2) = \frac{1}{N} \sum_a \text{var}(t_a) + \sigma_\tau^2 + \text{var}(\bar{l}) + 0 - 2 \text{var}(\bar{l}) - 0. \quad (10)$$

Now $\text{var}(t_a)$ is given by (2), so that

$$E(s_i^2) = \frac{1}{nN} \sum_a \tau_a(n - \tau_a) + \sigma_\tau^2 - \text{var}(\bar{l}). \quad (11)$$

Finally, proceeding as in (6), we have

$$E(s_i^2) = \bar{\tau} - \frac{1}{n} \bar{\tau}^2 + \frac{n-1}{n} \sigma_\tau^2 - \text{var}(\bar{l}). \quad (12)$$

Next, by the definition of $\text{var}(\bar{l})$,

$$E(\bar{l}^2) = \text{var}(\bar{l}) + \bar{\tau}^2. \quad (13)$$

From (7), (12), and (13),

$$\begin{aligned} &E[s_i^2(1 - r_{21})] \\ &= \frac{1}{n-1} \left[n\bar{\tau} - \text{var}(\bar{l}) - \bar{\tau}^2 - \bar{\tau} + \frac{1}{n} \bar{\tau}^2 - \frac{n-1}{n} \sigma_\tau^2 + \text{var}(\bar{l}) \right] \\ &= \bar{\tau} - \frac{1}{n} \sigma_\tau^2 - \frac{1}{n} \bar{\tau}^2. \end{aligned} \quad (14)$$

This result is the same as that in (6). We have shown that the average squared standard error of measurement found in type-2 sampling is exactly equal to the expected value of the squared S.E. Meas. derived from the formula-21 Kuder-Richardson reliability coefficient.

The logical relation between Kuder-Richardson formulas 20 and 21 can be derived from (1) and (5), from which it is readily found that

$$s_i^2(1 - r_{20}) = s_i^2(1 - r_{21}) - \frac{n^2}{n - 1} s_p^2. \quad (15)$$

Now the term on the left and the first term on the right of (15) are the squared standard errors of measurement computed from r_{20} and from r_{21} , respectively. Furthermore, since $ns_p^2/(n - 1)$ is the unbiased small-sample estimate of the population variance σ_p^2 , it is seen that the last term on the right is the small-sample estimator for the squared standard error of the mean score [see (22)]. Consequently, we may rewrite (15) as

$$(\text{S.E.Meas.}_{20})^2 = (\text{S.E.Meas.}_{21})^2 - \text{S.E.}^2(\bar{t}). \quad (16)$$

The difference between r_{20} and r_{21} , as made apparent in (16), arises from the fact that some randomly parallel forms are, by chance, composed of harder-than-average items, or of easier-than-average items; consequently, the mean of the actual scores on any given test is not exactly equal to the mean of the true scores for the same examinees. *The use of r_{20} is appropriate whenever one is willing to ignore any difference between the mean test score of the group and their mean true score, i.e., when one is concerned only with the relative rather than the absolute size of the scores of the group. On the other hand, r_{21} should be used whenever one is concerned with the actual magnitude of the errors of measurement, e.g., whenever there is a predetermined cutting score which divides the examinees into passing and failing groups.*

The foregoing treatment brings to our attention the very important fact that $\text{S.E.}(t_a)$ is actually the same as the traditional standard error of measurement of the individual examinee's score. The first formula in the second column of Table 1 thus provides a very simple way of computing this important quantity.

5. Comparison with Certain Standard Formulas

A formula closely related to (4) is the following, adapted from (66) of reference (8), which will appear familiar to most readers:

$$\text{S.E.}(\bar{t}) = \frac{s_t}{\sqrt{N}} \sqrt{1 - \text{reliability}}. \quad (66')$$

The question arises as to why $\text{S.E.}(\bar{t})$ in (66') has a totally different formula from that given in Table 1 for the type-2 standard error of the mean. If we use (66') to determine whether or not two forms of a test yield significantly different mean scores, we will always find the difference to be significant provided only that we take a sufficiently large number of examinees (N) for our experiment. This is true because the standard error in (66') is inversely proportional to \sqrt{N} —the standard error vanishes when N is large.

(66') represents the sampling fluctuations of the mean that would be observed if the same test were administered to successive samples of N examinees so chosen that the distribution of true scores was the same in each sample. If the same test is administered twice to the same group of examinees, (66') could be used in investigating the significance of the difference between the mean scores obtained on the two testings, provided it can be assumed that there is no practice effect. In this case, there is only one test involved, and there is thus no sampling of test items. Obviously, (66') should not be used when there is sampling of items—a type-2 standard error is required.

Consider next Wilks' (11) and Votaw's (10) procedures when either of these is used as a criterion of "parallelism" in tests, as suggested by Gulliksen (3, Ch. 14). Gulliksen defines "parallel" tests as having equal means, equal variances, and equal intercorrelations with each other and with all external criteria (as well as satisfying appropriate non-statistical criteria of parallelism). Wilks' and Votaw's significance tests provide rigorous statistical criteria for "parallelism" under this definition. They could appropriately be applied if identically the same tests were administered twice to the same examinees, provided it could be assumed that no practice effect had occurred. It would not be very desirable, however, to apply Wilks' or Votaw's procedures to data such as were obtained in the second illustrative example given in section 3. If a test composed of items having a certain characteristic is to be compared with a test composed of different items having a second characteristic, it may not be very useful to set up the null hypothesis that the two tests are strictly interchangeable in every way. Such a null hypothesis will always be rejected if N is sufficiently large, but the rejection of this hypothesis does not necessarily imply that the first and second characteristics have different effect, since the observed discrepancy might be readily accounted for as no greater than would be expected to be found in comparing two randomly parallel tests composed of the same kind of items.

6. *Sampling Distributions of Test Statistics*

It remains only to present the derivations of the results that have up to now been quoted without proof. The derivations are based on the assertion that there is a definite response (x_{ia}) that a given examinee will make to a given item. The nature of this response may or may not be known in advance. The group of N examinees to whom the items or tests are administered is a fixed group not subject to sampling fluctuation or other changes.

The responses of the N examinees to item i may be specified by the column vector $\{x_i = x_{i1}, x_{i2}, \dots, x_{iN}\}$. Since each item response is assumed to be treated as either "right" or "wrong," $x_{ia} = 0$ or 1, and there are exactly 2^N possible different vectors, i.e., different patterns of item response. If we let the subscript $I = 1, 2, 3, \dots, 2^N$, then these possible patterns are represented by the 2^N vectors x_I . If two items have exactly the same pattern of

responses, i.e., if the response of each examinee is the same on both items, then the two items are wholly indistinguishable in the present situation. It may therefore be asserted without loss of generality that, for present purposes, any infinite pool of items is composed of 2^N different kinds of items, designated by the 2^N vectors x_I . The relative frequencies of occurrence of the different kinds of items are therefore the only parameters needed to describe completely any infinite pool; these parameters will be denoted by π_I , the relative frequencies of occurrence of the patterns x_I .

When a random sample of n test items is drawn from the pool, the probability that the resulting n -item test will be composed of n_1 items of the first kind, n_2 items of the second kind, \dots , n_I items of the I th kind, \dots , $n_{(2^N)}$ items of the 2^N th kind is given by the standard multinomial distribution (7, pp. 58-59):

$$f(n_1, n_2, \dots, n_{(2^N)}) = \frac{n!}{\prod_I n_I!} \prod_I \pi_I^{n_I}. \quad (17)$$

It can be shown (1, p. 419) that the quantities $V_I = (n_I - n\pi_I)/\sqrt{n\pi_I}$ are asymptotically normally distributed for large n with zero means and with the (singular) variance-covariance matrix $I - \pi\pi'$, where I is the identity matrix and π is the column vector $(\sqrt{\pi_1}, \sqrt{\pi_2}, \dots, \sqrt{\pi_{(2^N)}})$. Now, the test score of individual a is $z_a = \sum_i x_{ia}/n = \sum_I x_{Ia}n_I/n$, the x_{Ia} being given constants, 0 or 1, not subject to sampling fluctuation; or, in terms of V_I ,

$$z_a = \sum_I \pi_I x_{Ia} + \frac{1}{\sqrt{n}} \sum_I \sqrt{\pi_I} x_{Ia} V_I.$$

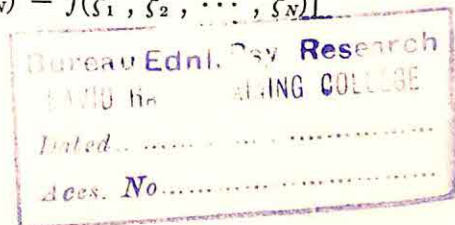
The first term on the right is $\zeta_a = \tau_a/n$, the "true" proportion-correct score; so that, finally,

$$\sqrt{n}(z_a - \zeta_a) = \sum_I \sqrt{\pi_I} x_{Ia} V_I.$$

It is thus seen that the N variables $\sqrt{n}(z_a - \zeta_a)$ are asymptotically jointly multinormally distributed, each with a mean of zero, a variance which turns out to be $\zeta_a(1 - \zeta_a)$, and covariances $\zeta_{ab} - \zeta_a\zeta_b$, where ζ_{ab} is the proportion of all items answered correctly by both examinee a and examinee b . It follows immediately that the large-sample standard error of z_a is $\sqrt{\zeta_a(1 - \zeta_a)/n}$ [cf. (2)]. The derivation of these and other standard errors will be left to the following section.

By a well-known theorem, if $f(z_1, z_2, \dots, z_N)$ is a function of the z_a having continuous first-order partial derivatives with respect to each z_a at the point $(\zeta_1, \zeta_2, \dots, \zeta_N)$, and if at least one of these derivatives is non-vanishing at this point, then the quantity

$$\sqrt{n}[f(z_1, z_2, \dots, z_N) - f(\zeta_1, \zeta_2, \dots, \zeta_N)]$$



is asymptotically normally distributed with zero mean when n is sufficiently large. This theorem assures us that *the mean score* (\bar{z} or \bar{l}), *the standard deviation of the scores* (s_z or s_l), *the Kuder-Richardson formula-21 reliability* (r_{21}), and *the test validity* (r_{cz} or r_{cl}), are approximately normally distributed in type-2 sampling with large n ; and in addition gives us the large-sample expected value of each statistic. It seems highly likely that the Kuder-Richardson reliability, formula 20, likewise is asymptotically normally distributed, but no proof of this conclusion is available at present, in view of the fact that the formula for this statistic involves $\sigma^2(p)$, which is not a function of the z_a .

The foregoing proof of asymptotic normality follows a line of reasoning that would require n to be very large except when N is very small, viz. $N = 2$. The nature of the situation, however, gives excellent reason to suppose that normality is approximated more quickly than the line of proof suggests when $N > 2$. No rigorous proof of this fact has been found.

7. Derivations of Expected Values and Standard Errors

The Individual Score

The proportion of the items in the entire pool to which examinee a will give the correct answer is, by definition, $\zeta_a = \tau_a/n$. If we concern ourselves with only a single examinee, the number of correct responses that he gives on one sample of items is *not* correlated with the number that he gives on other samples. If n items are drawn at random from the pool, t_a , the score of examinee a on the resulting test, i.e., the number of items that he will answer successfully, will of necessity have the usual binomial distribution with mean and variance

$$E(t_a) = \tau_a, \quad (18)$$

$$\text{S.E.}^2(t_a) = \frac{1}{n} \tau_a(n - \tau_a) = n\zeta_a(1 - \zeta_a). \quad (19)$$

This conclusion (and also those that follow, except as large n may be assumed) depends on no assumptions whatever except that of random sampling. (19) is identical with (2), which was discussed in a previous section. If the observed value t_a is substituted for the unknown τ_a in (19), we obtain the square of the first formula of Table 1.

For finite sampling, when n items are drawn without replacement from a finite pool of m items, the corresponding formulas, stated without proof, are

$$E(t_a) = \tau_a, \quad (18')$$

$$\text{S.E.}^2(t_a) = \frac{m-n}{mn} \tau_a(n - \tau_a). \quad (19')$$

The Mean Score of the Group Tested

It should be noted that the scores of examinees a and b are not independent over different parallel forms of the test. If a particular form happens to be composed of rather difficult items, both examinees will tend to get low scores; if a particular form happens to be easy, both will tend to score higher. Consequently, although the expected value of the mean score in the group is equal to the mean of the expected values of the individual scores, i.e.,

$$E(\bar{t}) = \frac{1}{N} \sum_a \tau_a = \bar{\tau}, \quad (20)$$

the standard error of the mean is not an average of the standard errors of the individual scores.

It will be convenient from this point on to work with $z_a = t_a/n$, the proportion-correct score, rather than with t_a itself. The nature of the desired standard error follows immediately from the fact that the mean score (\bar{z}) is identically equal to the average item difficulty

$$\bar{z} \equiv M(p). \quad (21)$$

The usual formulas for the standard error of a mean apply to $M(p)$, so that

$$\text{S.E.}^2(\bar{z}) = \frac{1}{n} \sigma^2(p), \quad (22)$$

where $\sigma(p)$ is the standard deviation of the item difficulties over the whole pool of items. If the observed value of $s^2(p)$ is substituted for the unknown $\sigma^2(p)$, and if t/n is substituted for z , the square of the second formula of Table 1 is obtained. [(19) is a special case of (22), being obtained when $p_i = x_{ia}$.]

In sampling from a finite pool of m items, the corresponding formula, stated without proof, is

$$\text{S.E.}^2(\bar{z}) = \frac{m-n}{mn} \sigma^2(p). \quad (22')$$

We may note that $\sigma(p)$ for a given set of items, and hence $\text{S.E.}_2(\bar{z})$ for a given test, will be higher when N is small than when N is large. Suppose, for example, that all items have the same difficulty (p) for a very large group of examinees, so that for this group $\sigma(p) = 0$. If the same items are administered to a smaller group of examinees drawn at random from the larger, the observed values of p_i in the smaller group will differ from each other because of type-1 sampling fluctuations, and $\sigma(p)$ will be greater than zero. In the extreme case where $N = 1$, the observed values of p are of necessity either 0 or 1, and $\sigma(p)$ is at a maximum.

The Standard Deviation of the Scores of the Group Tested

In order to obtain the standard error of s_z^2 , we first use the formula for the variance of a sum to write

$$s_z^2 = \frac{1}{n^2} \sum_h \sum_i s_{ih}, \quad (23)$$

s_{ih} being the covariance between item i and item h . Then, again from the formula for the variance of a sum,

$$\text{var}(s_z^2) = \frac{1}{n^4} \sum_h \sum_i \sum_j \sum_k \text{cov}(s_{ih}, s_{jk}), \quad (24)$$

where "cov" stands for the sampling covariance

$$\text{cov}(s_{ih}, s_{jk}) = E s_{ih} s_{jk} - E s_{ih} E s_{jk}.$$

Grouping the sums in (24), we obtain

$$\begin{aligned} \text{var } s_z^2 = \frac{1}{n^4} & \left[\sum_{(h \neq i \neq j \neq k)}^{(n^4 - 6n^3 + 11n^2 - 6n)} \text{cov}(s_{hi}, s_{jk}) + 2 \sum_{(i \neq j \neq k)}^{(n^3 - 3n^2 + 2n)} \text{cov}(s_i^2, s_{jk}) \right. \\ & + 4 \sum_{(i \neq j \neq k)}^{(n^3 - 3n^2 + 2n)} \text{cov}(s_{ii}, s_{jk}) + 4 \sum_{(i \neq j)}^{(n^2 - n)} \text{cov}(s_i^2, s_{ii}) \\ & \left. + \text{other sums containing no more than } n^2 \text{ terms each} \right]. \quad (25) \end{aligned}$$

Here the first sum is over all sets of four subscripts no two of which are the same, etc. The coefficient 2 of the second sum arises from combining the two equivalent expressions $\sum_{(i \neq j \neq k)} \text{cov}(s_i^2, s_{jk})$ and $\sum_{(h \neq i \neq j)} \text{cov}(s_{hi}, s_j^2)$. The other numerical coefficients arise similarly. The polynomials in n written above the summation signs indicate the number of terms involved in the summation.

Now, the terms under each set of summation signs in (25) are all the same no matter what the numerical values of the subscripts; consequently

$$\begin{aligned} \text{var } s_z^2 = \frac{1}{n^4} & [(n^4 - 6n^3 + 11n^2 - 6n) \text{cov}(s_{hi}, s_{jk}) \\ & + 2(n^3 - 3n^2 + 2n) \text{cov}(s_i^2, s_{jk}) \\ & + 4(n^3 - 3n^2 + 2n) \text{cov}(s_{ii}, s_{jk}) + O(n^2)], \quad (26) \end{aligned}$$

where $O(n^2)$ stands for terms of order n^2 . In (26) and in the following paragraph it is understood that $h \neq i \neq j \neq k$.

Now, s_{hi} and s_{jk} fluctuate independently over successive samples, so that $\text{cov}(s_{hi}, s_{jk}) = 0$. The same is true of s_i^2 and s_{jk} . Consequently,

$$\begin{aligned} \text{var } s_z^2 &= \frac{4}{n^4} (n^3 - 3n^2 + 2n) \text{cov}(s_{ij}, s_{ik}) \\ &+ O\left(\frac{1}{n^2}\right) = \frac{4}{n} \text{cov}(s_{ij}, s_{ik}) + O\left(\frac{1}{n^2}\right). \end{aligned} \quad (27)$$

Equation 27 gives the desired result, but not in a very useful form, since $\text{cov}(s_{ij}, s_{ik})$ is a function of population parameters and is generally not known. As a final step, then, it will be shown that $s^2(s_{iz})$, the actual variance (over items 1 to n) of the observed item-test covariances, provides a consistent estimate of $\text{cov}(s_{ij}, s_{ik})$; it will be proved that

$$E[s^2(s_{iz})] = \text{cov}(s_{ij}, s_{ik}) + O\left(\frac{1}{n}\right). \quad (28)$$

From the formula for the covariance of a sum,

$$s_{iz} = \frac{1}{n} \sum_j s_{ij}; \quad (29)$$

$$s^2(s_{iz}) = \frac{1}{n^2} \sum_i \sum_k s(s_{ij}, s_{ik}), \quad (30)$$

the term under the summation sign being the actual covariance (over items 1 to n) of the observed values of s_{ij} and s_{ik} :

$$s(s_{ij}, s_{ik}) = \frac{1}{n} \sum_i s_{ij}s_{ik} - \frac{1}{n^2} \left(\sum_i s_{ij} \right) \left(\sum_i s_{ik} \right). \quad (31)$$

Substituting from (31) into (30), and taking expected values, we find

$$E[s^2(s_{iz})] = \frac{1}{n^3} \sum_i \sum_j \sum_k E(s_{ij}s_{ik}) - \frac{1}{n^4} \sum_h \sum_i \sum_j \sum_k E(s_{hi}s_{ik}). \quad (32)$$

Grouping the sums on the right, we have

$$\begin{aligned} E[s^2(s_{iz})] &= \frac{1}{n^3} \left[\sum_{(i \neq j \neq k)}^{n(n-1)(n-2)} E(s_{ij}s_{ik}) + O(n^2) \right] \\ &\quad - \frac{1}{n^4} \left[\sum_{(h \neq i \neq j \neq k)}^{n(n-1)(n-2)(n-3)} E(s_{hi}s_{ik}) + O(n^3) \right]. \end{aligned} \quad (33)$$

Now, the terms under each summation sign in (33) are the same regardless of the numerical value of the subscript. Furthermore, as already pointed out in deriving (27), $\text{cov}(s_{hi}, s_{ik}) = 0$ when $h \neq i \neq j \neq k$, or in other words, $E(s_{hi}s_{ik}) - E(s_{hi})E(s_{ik}) = 0$, or $E(s_{hi}s_{ik}) = E(s_{ij})E(s_{ik})$. Consequently,

$$E[s^2(s_{iz})] = E(s_{ij}s_{ik}) - E(s_{ij})E(s_{ik}) + O\left(\frac{1}{n}\right). \quad (34)$$

But this is the same as (28), which was to be proved.

The large sample standard error of s_z^2 may therefore be estimated from the actual variance of the observed item-test covariances:

$$\text{S.E.}^2(s_z^2) = \frac{4}{n} s^2(s_{iz}). \quad (35)$$

By means of the "delta" method (5, Vol. 1, pp. 208 ff.), it is readily shown from (35) that in large samples

$$\text{S.E.}^2(s_z) = \frac{1}{4s_z^2} \text{S.E.}^2(s_z^2) = \frac{s^2(s_{iz})}{ns_z^2}. \quad (36)$$

If t/n is substituted for z in (36), the square of the third equation of Table 1 is obtained.

The corresponding squared standard error for sampling from finite populations may be shown to be

$$\text{S.E.}^2(s_z^2) = 4 \frac{m-n}{mn} s^2(s_{iz}). \quad (37)$$

The Kuder-Richardson Reliability Coefficient, Formula 20

Let the usual formula for r_{20} , the Kuder-Richardson formula-20 coefficient, be rewritten as follows:

$$r_{20} = \frac{n}{n-1} \left(1 - \frac{R}{n} \right), \quad (38)$$

where

$$R = \frac{1}{n} \sum_i s_i^2 / s_z^2 = M / s_z^2, \quad \text{say.}$$

In the extraordinary case where $s_z^2 = 0$, we will agree not to try to compute any value of r_{20} . The "delta" method may now be used to obtain the result

$$\text{var } R = \frac{1}{s_z^4} \text{var } M + \frac{M^2}{s_z^8} \text{var } s_z^2 - \frac{2M}{s_z^6} \text{cov}(M, s_z^2). \quad (39)$$

Now $\text{var}(s_z^2)$ is already known from (35). $\text{Var}(M)$ can be evaluated by the usual formula for the standard error of a mean:

$$\text{var } M = \frac{1}{n} s^2(s_i^2), \quad (40)$$

where $s^2(s_i^2)$ is the actual variance of the observed item variances. Finally, it is readily shown, by methods similar to those used in evaluating $\text{var}(s_z^2)$, that

$$\text{cov}(M, s_z^2) = \frac{2}{n} s(s_i^2, s_{iz}), \quad (41)$$

where $s(s_i^2, s_{iz})$ is the actual covariance between the observed item variances and the observed item-test covariances. Consequently,

$$\text{var } R \doteq \frac{1}{ns_z^4} [s^2(s_i^2) + 4R^2 s^2(s_{iz}) - 4Rs(s_i^2, s_{iz})]. \quad (42)$$

Now $\text{var}(r_{20}) = \text{var}(R)/n^2$; hence, to order $1/n^4$,

$$\text{S.E.}^2(r_{20}) = \frac{1}{n^3 s_z^4} [s^2(s_i^2) + 4n^2(1 - r_{20})^2 s^2(s_{iz}) - 4n(1 - r_{20})s(s_i^2, s_{iz})]. \quad (43)$$

It may be noted that the quantity $(1 - r_{20})$ is of order $1/n$, because $\lim_{n \rightarrow \infty} n(1 - r_{20}) = \text{constant}$. It is then seen from (43) that $\text{S.E.}^2(r_{20})$ is a quantity of order $1/n^3$. Equation 43 leads directly to the fourth formula of Table 1.

It may be shown that the corresponding standard error when sampling from a finite population is $(m - n)/m$ times the value given in (43).

The Kuder-Richardson Reliability Coefficient, Formula 21

By a procedure wholly parallel to that used for the formula-20 reliability coefficient, it is found that, approximately,

$$\begin{aligned} \text{S.E.}^2(r_{21}) = \frac{1}{n^3 s_z^4} [(1 - 2\bar{z})^2 s^2(p) + 4n^2(1 - r_{21})^2 s^2(s_{iz}) \\ - 4n(1 - r_{21})(1 - 2\bar{z})s(p_i, s_{iz})], \end{aligned} \quad (44)$$

where $s(p_i, s_{iz})$ is the actual covariance between the observed item difficulties and the observed item-test covariances. Equation (44) leads directly to the fifth formula of Table 1.

The standard error of the split-half reliability coefficient has not been worked out. It must, however, be larger than the standard error of r_{20} , given by (43), since r_{20} is the mean of the split-half coefficients from all possible splits, as shown by Cronbach (2).

The Validity Coefficient

If c is an outside criterion,

$$r_{cz} = \frac{s_{cz}}{s_c s_z}. \quad (45)$$

By the "delta" method,

$$\text{var } r_{cz} = r_{cz}^2 \left[\frac{\text{var } s_{cz}}{s_{cz}^2} + \frac{\text{var } s_z^2}{4s_z^4} - \frac{\text{cov}(s_{cz}, s_z^2)}{s_{cz} s_z^2} \right]. \quad (46)$$

It is found that

$$\text{var } s_{cz} \doteq \frac{1}{n} s^2(s_{ci}); \quad (47)$$

$$\text{cov}(s_{iz}, s_z^2) \doteq \frac{2}{n} s(s_{ci}, s_{iz}). \quad (48)$$

Finally,

$$\text{S.E.}^2(r_{cz}) = \frac{1}{ns_z^2} \left[\frac{1}{s_c^2} s^2(s_{ci}) - \frac{2r_{cz}}{s_z s_c} s(s_{ic}, s_{iz}) + \frac{r_{cz}^2}{s_z^2} s^2(s_{iz}) \right]. \quad (49)$$

Equation (49) leads directly to the last formula of Table 1.

The corresponding standard error for sampling from a finite pool of items is presumably $(m - n)/m$ times the foregoing quantity.

8. *Simultaneous Sampling of Items and Examinees*

Simultaneous and independent sampling of items and examinees might be called matrix sampling instead of type-12 sampling. [A generalized approach to this problem is reported in (4).] Here, both the population and the sample may be thought of as matrices. Each row of the population matrix may be taken as representing one test item, and each column as representing one examinee. The elements of the matrix are taken to be 1's and 0's, depending upon whether or not the examinee would answer the item correctly if it were administered to him. The actual responses given by a random sample of examinees to a test consisting of a random sample of items can be thought of as constituting a rectangular matrix composed of n rows and N columns selected independently and at random from the population matrix.

Let y be any statistic calculated from the sample matrix. Consider all possible $n \times N$ matrices that can be formed from the population matrix by a process of omitting entire rows and columns. $\text{Var}_{12} y$, the type-12 sampling variance of y , is, by definition, equal to the variance of the y values calculated from all possible such $n \times N$ matrices, i.e.,

$$\text{var}_{12} y = E_{12}(y - E_{12}y)^2, \quad (50)$$

where E_{12} indicates that the expectation of the directly following quantity is to be taken over all possible $n \times N$ matrices. (The convention of always following each expectation symbol with parentheses or brackets will be dropped.)

Equation (50) may be made more convenient by application of a very familiar lemma from analysis of variance, which states that the "total sum of squares" is equal to the "within sum of squares" plus the "among sum of squares." It is immediately found that

$$\text{var}_{12} y = E_1[E_{2.1}(y - E_{2.1}y)^2] + E_1(E_{2.1}y - E_{12}y)^2, \quad (51)$$

where $E_{2.1}$ is the conditional expected value over all possible combinations of rows of the population matrix, the columns being held fixed, and E_1 is the expected value over all possible combinations of columns. In more concise notation, (51) becomes

$$\text{var}_{12} y = E_1(\text{var}_{2.1} y) + \text{var}_1(E_{2.1}y) \quad (52)$$

where $\text{var}_{2.1}$ and var_1 are type-2 and type-1 sampling variances, respectively. By symmetry, there is also the alternative equation

$$\text{var}_{12} y = E_2(\text{var}_{1.2} y) + \text{var}_2 (E_{1.2} y). \quad (53)$$

If y is a consistent statistic in type-2 sampling, $E_{2.1}y$ will not differ greatly from y in large samples. This fact suggests that it will often be found in large samples that

$$\text{var}_1 (E_{2.1} y) \doteq \text{var}_1 y. \quad (54)$$

Similarly

$$E_1(\text{var}_{2.1} y) \doteq \text{var}_2 y. \quad (55)$$

If (54) and (55) hold to a satisfactory order of approximation, then (52) reduces to the very simple result that the type-12 sampling variance is approximately equal to the sum of the type-1 and type-2 sampling variances

$$\text{var}_{12} y \doteq \text{var}_1 y + \text{var}_2 y. \quad (56)$$

A similar statement can be made for (53).

The simple result represented by (56) can be shown to hold in the case of the mean score, \bar{z} or \bar{l} , and, at least under the assumption that the scores are normally distributed, in the case of the standard deviation, s_z or s_l . Proofs are presented in the following two sections.

The type-1 sampling variances of the Kuder-Richardson reliability coefficients are not known to the writer. Since the type-2 sampling variances of r_{20} and r_{21} are of order $1/n^3$ [see (43) and (44)], it seems clear that the type-12 sampling variances of these coefficients, to our order of approximation, depend only on the unknown type-1 sampling variances. Neither these nor the type-12 sampling variances of the reliability or validity coefficients have been worked out.

The Mean Score

In the case of the mean relative test score, from (20), (22), and (52),

$$\text{var}_{12} (\bar{z}) = \frac{1}{n} E_1 \sigma^2(p) + \text{var}_1 \bar{\xi}, \quad (57)$$

where $\sigma^2(p)$ is the variance over all items in the population of the values of p_i for a given group of examinees, and $\bar{\xi} = \sum_a \xi_a / N$.

According to the standard formula for the standard error of any mean, the last term of (57) is

$$\text{var}_1 \bar{\xi} = \frac{1}{N} \sigma_{\xi}^2, \quad (58)$$

where σ_{ξ}^2 is the standard deviation of ξ_a over the entire population of examinees.

Next, it will be helpful to evaluate the first term on the left of (57)

$$\frac{1}{n} E_1 \sigma_{2.1}^2(p) = \frac{1}{n} E_1 (E_{2.1} p_i^2 - \bar{\xi}^2) = \frac{1}{n} (E_2 E_1 p_i^2 - E_1 \bar{\xi}^2). \quad (59)$$

Now the difference $E_1 p_i^2 - \pi_i^2$, where $\pi_i = E_1 p_i$, is by definition $\text{var}_1 p_i$, the usual binomial variance known to equal $\pi_i(1 - \pi_i)/N$. Hence,

$$E_1 p_i^2 = \frac{N-1}{N} \pi_i^2 + \frac{\pi_i}{N}. \quad (60)$$

Similarly,

$$E_1 \bar{\xi}^2 = \frac{1}{N} \sigma_\pi^2 + \bar{Z}^2, \quad (61)$$

where $\bar{Z} = E_1 \bar{\xi}_a = E_2 \pi_i = E_{12} z_a = E_{12} p_i = E_{12} x_{ia}$ is the over-all population mean. From (60),

$$E_2 E_1 p_i^2 = \frac{N-1}{N} E_2 \pi_i^2 + \frac{1}{N} \bar{Z}. \quad (62)$$

As before,

$$E_2 \pi_i^2 = \sigma_\pi^2 + \bar{Z}^2, \quad (63)$$

where σ_π^2 is the standard deviation of the values of π_i over the entire population of items. From (62) and (63),

$$E_2 E_1 p_i^2 = \frac{N-1}{N} \sigma_\pi^2 + \frac{N-1}{N} \bar{Z}^2 + \frac{1}{N} \bar{Z}. \quad (64)$$

The substitution first of (58) and (59), then of (61) and (64) into (57) gives finally

$$\text{var}_{12}(\bar{z}) = \frac{1}{nN} [(n-1)\sigma_\pi^2 + (N-1)\sigma_\pi^2 + \bar{Z}(1-\bar{Z})]. \quad (65)$$

Equation (65) gives the exact, small-sample sampling variance of \bar{z} . The same result can be obtained from (53), as a check.

When terms of order $1/n^2$ and of order $1/N^2$ are neglected in (65) the large-sample type-12 sampling variance is found to be approximately

$$\text{var}_{12}(\bar{z}) \doteq \frac{1}{N} \sigma_\pi^2 + \frac{1}{n} \sigma_\pi^2. \quad (66)$$

Since $\sigma_\pi^2 = E_2 s_z^2$ [see (70)], it is easily seen that to our order of approximation

$$\text{S.E.}_{12}(\bar{z}) = \frac{1}{N} s_z^2 + \frac{1}{n} s_p^2. \quad (67)$$

We thus have the simple result that the type-12 sampling variance of the mean test score is equal to the sum of the type-1 and the type-2 sampling variances approximately.

The Standard Deviation of Scores

In the case of s_z^2 we find by dividing (12) by n^2 that

$$E_2 s_z^2 = \frac{\bar{z}(1 - \bar{z})}{n} + \frac{n-1}{n} \sigma_z^2 - \text{var } \bar{z}; \quad (68)$$

or, dropping terms of order $1/n$,

$$E_2 s_z^2 \doteq \sigma_z^2, \quad (69)$$

as might be expected.

Also, a standard formula gives the result

$$\text{var}_1 s_z^2 = \frac{1}{N} [\mu_4(z) - \sigma_z^4], \quad (70)$$

where $\mu_4(z)$ and σ_z^4 are the fourth and the squared second moments of the distribution of the scores of all examinees. If we are willing, for the sake of simplicity, to assume that these scores are effectually normally distributed, then

$$\text{var}_1 s_z^2 = \frac{2}{N} \sigma_z^4. \quad (71)$$

From (71), (69), and (53), approximately,

$$\text{var}_{12} s_z^2 \doteq \frac{2}{N} E_2 \sigma_z^4 + \text{var}_2 \sigma_z^2. \quad (72)$$

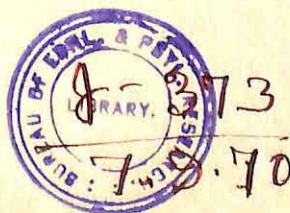
From (72), (69), and (35),

$$\text{var}_{12} s_z^2 \doteq \frac{2}{N} \sigma_z^4 + \frac{4}{n} s^2(\sigma_{iz}), \quad (73)$$

where σ_z is the standard deviation of all true scores and $s^2(\sigma_{iz})$ is the variance over n items of the true item-test covariances computed using all examinees in the population. To our order of approximation,

$$\text{S.E.}_{12}(s_z^2) = \frac{2}{N} s_z^4 + \frac{4}{n} s^2(s_{iz}). \quad (74)$$

Under the assumption that z is effectually normally distributed, it is thus found that the type-12 sampling variance of s_z^2 is approximately equal to the sum of the type-1 and type-2 sampling variances.



REFERENCES

1. Cramér, H. Mathematical methods of statistics. Princeton Univ. Press, 1946.
2. Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, 16, 297-334.
3. Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.
4. Hooke, R. Sampling from a matrix, with applications to the theory of testing. Statistical Research Group, Princeton University. Memorandum Report 53, 1953. (Dittoed.)
5. Kendall, M. G. The advanced theory of statistics. London: Charles Griffin and Co., 1948. 2 vols.
6. Kuder, G. F. and Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika*, 1937, 2, 151-160.
7. Mood, A. M. Introduction to the theory of statistics. New York: McGraw-Hill, 1950.
8. Peters, C. C. and Van Voorhis, W. R. Statistical procedures and their mathematical bases. New York: McGraw-Hill, 1940.
9. Tucker, L. R. A note on the estimation of test reliability by the Kuder-Richardson formula (20). *Psychometrika*, 1949, 14, 117-119.
10. Votaw, D. F., Jr. Testing compound symmetry in a normal multivariate distribution. *Ann. math. Stat.*, 1948, 19, 447-473.
11. Wilks, S. S. Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. *Ann. math. Stat.*, 1946, 17, 257-281.

Manuscript received 3/1/54

Revised manuscript received 5/11/54

SEPARATION OF DATA AS A PRINCIPLE IN FACTOR ANALYSIS

CHESTER W. HARRIS

UNIVERSITY OF WISCONSIN

Two systems of factor analysis—factoring correlations with units in the diagonal cells and factoring correlations with communalities in the diagonal cells—are considered in relation to the commonly used statistical procedure of separating a set of data (scores) into two or more parts. It is shown that both systems of factor analysis imply the separation of the observed data into two orthogonal parts. The matrices used to achieve the separation differ for the two systems of factor analysis.

One of the recurring operations in statistical work is that of separating data into parts. Probably the most common example of this is that of separating the raw score for each of a number of subjects into a deviation score plus the mean of the scores for these subjects. The analysis of variance offers examples of this practice, since this method of analysis in effect further separates such deviation scores into two or more parts, depending upon the complexity of the design. Similarly, linear regression theory postulates a separation of the data of the dependent variable into two parts and provides a method of calculating each. It is at least intuitively evident that factor analysis also implies a separation of data into parts; however, the particular characteristics of the principle followed in making the separation may not be well understood. The purpose of this discussion is to interpret the procedures of factor analysis from this point of view.

Cochran's theorem (3), particularly Cramér's discussion of it (4, pp. 116-18), shows the necessary and sufficient conditions for decomposing a sum of squares into orthogonal parts. Consider only the matrices of the quadratic forms. For such a decomposition, these matrices satisfy the equation

$$I = A_1 + A_2 + \cdots + A_m, \quad (1)$$

with the A_i symmetric idempotent matrices that are pairwise orthogonal and whose ranks sum to the rank of I . Idempotent matrices are singular matrices such that $A = A^2$; they may be viewed as being generated by incomplete orthogonal matrices, i.e., sets of orthogonal columns. Thus (1) implies an orthogonal transformation. The most familiar example of this is the decomposition

$$\sum x_i^2 = n\bar{x}^2 + \sum (x_i - \bar{x})^2,$$

where the summation is over n measures. This might be written

$$XIX' = XA_1X' + XA_2X' \quad (2)$$

with X designating a single row vector of data and the A_1 and A_2 properly defined symmetric idempotent matrices. In this case the transformation is accomplished by an orthogonal matrix with each element of the first column consisting of the positive reciprocal of the square root of n . The matrix A_1 therefore is square, of order n , with each element the positive reciprocal of n . A_2 is square and symmetric, with each diagonal element equal to $(n - 1)/n$, and each off-diagonal element equal to the negative reciprocal of n . Aitken (1) demonstrates the independence of these forms for samples of normally distributed variables. The analysis of variance for a single variable implies the decomposition of the matrix A_2 in (2) into two or more pairwise orthogonal idempotent matrices. For the simplest design, A_2 is separated into two parts, one associated with the notion of "variance between" and the other with the notion of "variance within." The further separation of the matrix associated with "variance between" implies more complicated designs, such as a factorial arrangement of groups. Aitken's paper also gives necessary and sufficient conditions for the independence of two quadratic forms.

Since the symmetric idempotent matrices detailed in (1) are pairwise orthogonal, i.e., $A_iA_j = A_jA_i = 0$, it follows that for any matrix X , $(XA_i)(XA_j)' = (XA_j)(XA_i)' = 0$. This is a function of the matrices of the quadratic forms, and does not imply a particular distribution for the population from which X is drawn. Cochran's theorem shows that the sampling distribution of terms such as those on the right of (2) is known if X is a random sample from a normal population (univariate case). Bartlett (2) has discussed tests of significance for a decomposition of the form

$$X = XA_i + XA_j, \quad (3)$$

where X is a sample from a multi-variate normal population with mean zero. Here, A_i and A_j are two parts of the matrix A_2 as it was defined for equation (2). Two points have been emphasized. One is that choosing symmetric idempotent matrices that are pairwise orthogonal as the matrices of the quadratic forms gives a decomposition, as in (3), for which $(XA_i)(XA_j)' = (XA_j)(XA_i)' = 0$. Second, under certain assumptions regarding the nature of the data, sampling distributions of statistics derived from the parts given by such a decomposition are known. The remainder of the paper will consist of a discussion of the principle given by this first point in relation to factor analysis.

The following geometric representation of factor analysis is well known. For convenience, it will be assumed that the data have been scaled to unit variance; this may be a critical assumption from the statistical point of view, as Rao (8) shows, and the making of it emphasizes the attempt in this paper

to describe factor analysis rather generally and not to treat the complicated inferential problems. It is possible to regard the n persons as defining a space within which are located the k tests. The person axes are assumed to constitute a rectangular Cartesian system. Then the n measures for a given variable, when put in deviation form and scaled to unit variance, are the coordinates of a point in this person space that, when joined to the origin, defines the variable or test as a vector. Any factor may also be viewed as a unit-length vector located in this person space; such a factor is uniquely located by a set of n coordinates defining its end-point or by a set of direction cosines with respect to the n person axes. Define Z as a $k \times n$ matrix of data, such that $ZZ' = R_1$. The matrix of intercorrelations with units in the diagonals is designated by R_1 and Z' is the conventional transpose of Z . Let y be a column of direction cosines locating a single factor in the person space. Then Zy is the column of k scalar products of variables with this factor; these are correlation coefficients here, and may be regarded as a column of the factor matrix. Finally, $Z(yy')$ gives the coordinates, with respect to the person axes, of the perpendicular projections onto the factor axis of the points representing the variables. This expression $Z(yy')$ also is the portion of the data, Z , that is accounted for by the first factor. In other words,

$$Z = Z(yy') + Z(I - yy') \quad (4)$$

describes the separation of Z into two parts, one of which is associated with the factor that is located in the person space by the column of direction cosines, y , and the other part a remainder.

Equation (4) necessarily represents a separation of Z into two orthogonal parts. This is true because yy' is idempotent, i.e., $yy'yy' = yy'$, and consequently $I - yy'$ also is idempotent; therefore $yy'(I - yy') = (I - yy')yy' = 0$. It also is true that the matrix $Z(yy')$ is the least-squares approximation of the row y' to the rows of Z . This follows from least-squares theory; for a summary of the role of symmetrical idempotent matrices in generating least squares approximations see Harris (6). In general, then, the specification of y , i.e., the direction cosines of a single factor, provides a separation of the data Z into two orthogonal parts, one of which is the least-squares approximation of y' to Z . The only requirement imposed upon y has been that it designate a factor axis in the person space; in other words, y has been chosen arbitrarily from the indefinitely many possible unit-length vectors in the person space.

Equation (4) might be written more generally as

$$Z = ZA + Z(I - A), \quad (5)$$

where A designates a symmetrical idempotent matrix, i.e., $A = A^2$. Every symmetrical idempotent matrix may be viewed as the product YY' , where Y is a set of orthogonal columns, i.e., an incomplete orthogonal matrix. (If Y

is the complete orthogonal matrix, $A = I$, of course.) Equation (5) then is the case of selecting one or more mutually orthogonal unit-length axes in the person space as a set of factors; the direction cosines of these factors are given by Y . As before, the two parts on the right of (5) are uncorrelated and ZA is the least-squares approximation of Y' to Z . The matrix Y' is, of course, also regarded as the set of uncorrelated factor scores, each with unit variance. Again it should be emphasized that Y is arbitrary and might designate any set of orthogonal axes in the person space.

So far, then, it has been shown that the specification of one or more factors leads to the separation of Z into two orthogonal parts, one of which is a particular least-squares approximation.

The final step is to consider two approaches to factor analysis that differ primarily in the way in which the factor space is defined. The nature of these two approaches can be illustrated by considering the correlation between two variables. If the two variables are viewed as two unit-length vectors located in the person space, then the variable space is (at most) a plane, i.e., of dimension two. It is possible to define the factor space as identical with the variable space; this definition corresponds to choosing to factor the unit variances and the intercorrelations of the variables. If the complete intercorrelation matrix is non-singular, two factors may be extracted by this procedure. They would, necessarily, define the variable space. If only one factor is extracted, it would be represented by a line embedded in this variable space. Spearman's approach to this problem differs. His approach defines the common-factor space as the line formed by the intersection of two orthogonal planes, in each of which lies one of the unit-length vectors. The uncorrelated unique factors are defined by lines perpendicular to this single common-factor axis and, of course, also lie in these two intersecting planes. For two variables that are not correlated perfectly, i.e., are not collinear, the Spearman approach necessarily defines the common-factor space as *distinct* from the variable space. This definition corresponds to choosing to factor communalities and intercorrelations, rather than the correlation matrix with units in the diagonals.

The first approach, i.e., factoring R_1 , requires that any factor axis be embedded in the variable space; as a result, the factor might be located by reference to the k axes of the Cartesian system provided by the test vectors, as well as by reference to the n person axes. Obviously, the test vectors need not form a rectangular reference system. This means, then, that for any such factor there is some set of weights that, when applied to the variables, gives a linear combination of the entries in Z that reproduce the set of factor scores. Holzinger (7) gives illustrations of this principle, using both the centroid and what has since become known as the multiple-group methods of factor analysis. Using this approach, it may be pertinent to determine a "best" location of a factor. Eckart and Young (5) have shown the nature of

the best approximation, in a least-squares sense, of a matrix of data, Z , by another matrix of specified lower rank. Securing this best approximation is equivalent to identifying the first r principal-axis factors of R_1 , where r is the specified rank that is lower than the rank of Z . Defining the total factor space as identical with the total test space and then extracting r principal-axis factors from R_1 gives a separation of Z into the two parts of (5) such that $Z(I - A)Z'$ has a minimum trace compared with its trace for any other definition of A . That A is well-defined is evident from noting that A is generated by the r unit-length characteristic vectors of $Z'Z$ that correspond to the r largest characteristic roots of ZZ' , which necessarily are the same as those of $Z'Z$. The Eckart and Young results therefore show that their choice of Y gives a matrix ZA which is not only the least-squares approximation of Y' to Z , as it must be when A is generated by Y , but also a best approximation to Z .

Finally, it is evident that the communality principle in factor analysis also postulates an equation of the form of (5), since the common factors are defined by some set of direction cosines, Y . However, the common factors are not embedded in the variable space and consequently the elements of Y cannot be calculated from the data, Z . Thomson (9, p. 78) comments on this point. This means, then, that when equation (5) is used to describe the communality principle in factor analysis it must be regarded as a formal equation with $A = YY'$ unknown. If A were known, then a principal-axis resolution of ZA into factors and factor scores,

$$ZA = FS', \quad (6)$$

would lead to the definition of A as SS' . This would follow from noting that SS' is a unit for multiplication on the right of ZA that is of the same rank as A and recalling that a multiplication unit is unique within a group of singular matrices. However, this definition is circular, in that S , the factor matrix of factor scores, necessarily is identical with Y ; the unknown A remains unknown.

This discussion has emphasized the connection between factor analysis and well-known procedures for separating data into two or more parts. Following Cochran and Cramér, the separation of data into orthogonal parts was formulated in terms of symmetric idempotent matrices as the matrices of the quadratic forms. It was then shown that from the geometric view of factor analysis the specification of one or more factors is the specification of one or more sets of direction cosines that generate a symmetric idempotent matrix and that this matrix, A , and its annihilator, $(I - A)$, achieve a separation of the data. The nature of the matrix A was examined for two different approaches to factor analysis. For the first approach, Eckart and Young's results were reviewed to show that a minimum trace of $Z(I - A)Z'$ is achieved by the principal-axis factoring of R_1 . For the communality

approach, the merely formal character of A was emphasized. Although problems of estimation and statistical inference were not considered in this paper, this final result lends support to the belief that the communality principle poses important problems of statistical estimation.

REFERENCES

1. Aitken, A. C. On the independence of linear and quadratic forms in samples of normally distributed variables. *Proc. royal Soc. Edinburgh*, 1939, 60, 40-46.
2. Bartlett, M. S. Multivariate analysis. *J. royal stat. Soc. Sup*, 1947, 9, 176-90.
3. Cochran, W. G. The distribution of quadratic forms in a normal system with applications to the analysis of variance. *Proc. Cambridge phil. Soc.*, 1934, 30, 178-91.
4. Cramér, Harald. Mathematical methods of statistics. Princeton, N.J.: Princeton Univ. Press, 1946.
5. Eckart, Carl, and Young, Gale. The approximation of one matrix by another of lower rank. *Psychometrika*, 1936, 1, 211-18.
6. Harris, Chester W. The symmetrical idempotent matrix in factor analysis. *J. exp. Educ.*, 1951, 19, 239-46.
7. Holzinger, Karl J. Factoring test scores and implications for the method of averages. *Psychometrika*, 1944, 9, 155-67.
8. Rao, C. R. Estimation and tests of significance in factor analysis. (mimeographed).
9. Thomson, Godfrey. The factorial analysis of human ability. Boston: Houghton Mifflin Co., 1950. 4th edition.

Manuscript received 1/25/54

Revised manuscript received 4/9/54

THE CHOICE OF AN ERROR TERM IN ANALYSIS OF VARIANCE DESIGNS*

ARNOLD BINDER
INDIANA UNIVERSITY†

This article presents a survey of the assumptions which may be made in variance designs, a description of the mathematical models which reflect these assumptions, and a discussion of the ways in which various experimental conditions affect the choice of an error mean square. Particular emphasis is laid upon the principles, purposes, and dangers of pooling error mean squares in order to raise the power of a test. Specific recommendations are made for the rules of procedure for pooling (under various conditions) which produce tests with optimum power and error characteristics.

Among the various treatments of psychological statistics one finds a good deal of confusion and discrepancy in the recommended procedures for selecting an error term in the analysis of variance (see 4, 5, 7, as examples). In all too many cases the obtained significance or insignificance of the experimental results depends as much upon the particular statistics text used as upon the sampling data. The aim of this paper is to show the possible assumptions which may be made in regard to analysis of variance data, some of the hypotheses which may be tested, and how these and other factors influence the choice of the error term. Because of space limitations, the arguments will be restricted to a two-factor (or double classification) arrangement with m replications per cell. Many of the arguments presented here are directly translatable into the more complex designs.

Unfortunately, the derivations of the proper terms for testing various hypotheses under the conditions specified by the assumptions require a good deal of mathematical sophistication for their understanding. While references will, in all cases, be made to the sources in which the proofs may be found, this paper is aimed principally at the reader who is less interested in rigorous mathematical analysis than in the uses of the material in research design. For the purpose of identifying the various potential groups of assumptions and demonstrating the proper statistics under these assumptions, we shall make use of three mathematical models: linear hypothesis model, components of variance model, and mixed model.

*The writer is indebted to Professors Quinn McNemar and Lincoln Moses of Stanford University for reading the manuscript and offering many helpful suggestions and criticisms. He is grateful to Professor Z. W. Birnbaum of the University of Washington for preliminary suggestions as to form and notation.

†The preliminary draft of this paper was completed while the author was at Stanford University and the Veterans Administration Hospital, Palo Alto.

TABLE 1
Variance Schema for Two-Factor, m Replications Design

Source	D. of F.	Mean Square
Rows	$(r - 1)$	$\frac{cm \sum_{i=1}^r (X_{i..} - X_{...})^2}{(r - 1)} = s_r^2$
Columns	$(c - 1)$	$\frac{rm \sum_{j=1}^c (X_{.j.} - X_{...})^2}{(c - 1)} = s_c^2$
Interaction	$(r - 1)(c - 1)$	$\frac{m \sum_{i=1}^r \sum_{j=1}^c (X_{ij.} - X_{i..} - X_{.j.} + X_{...})^2}{(r - 1)(c - 1)} = s_i^2$
Within cells	$rc(m - 1)$	$\frac{\sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^m (X_{ijk} - X_{ij.})^2}{rc(m - 1)} = s_w^2$
Interaction plus Within cells	$(r - 1)(c - 1) + rc(m - 1)$	$\frac{m \sum_{i=1}^r \sum_{j=1}^c (X_{ij.} - X_{i..} - X_{.j.} + X_{...})^2 + \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^m (X_{ijk} - X_{ij.})^2}{(r - 1)(c - 1) + rc(m - 1)} = s_{i+w}^2$
Total	$(rcm - 1)$	$\frac{\sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^m (X_{ijk} - X_{...})^2}{(rcm - 1)}$

The data and definitions of Table 1 will be used throughout the paper.

In the table, X_{ijk} is the observation in the i th row ($i = 1, \dots, r$) and the j th column ($j = 1, \dots, c$) and for the k th replication ($k = 1, \dots, m$); and

$$X_{ij.} = \text{Observed cell mean} = \frac{\sum_{k=1}^m X_{ijk}}{m}$$

$$X_{i..} = \text{Observed row mean} = \frac{\sum_{j=1}^c \sum_{k=1}^m X_{ijk}}{cm}$$

$$X_{.j.} = \text{Observed column mean} = \frac{\sum_{i=1}^r \sum_{k=1}^m X_{ijk}}{rm}$$

$$X_{...} = \text{Observed over-all mean} = \frac{\sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^m X_{ijk}}{rcm}$$

Before dealing with the differences among the various models, let us consider the meaning of row, column, and interaction "effects." Each factor (or classification) is a characteristic or variable (such as individuals, conditions, tests, or treatments) which includes a number of different specific elements. In our case there are r elements in the data for the row factor and c elements in the data for the column factor. It is assumed in the analysis of variance that the value of each X_{ijk} observation is derived from two contributing sources: one dependent upon the particular row and column elements to which the particular unit belongs, the other independent of these elements. The first of the two contributing sources includes the row, column, and interaction effects; the second includes errors of observation and an over-all value constant for all of the observations in the data. Thus, the row effect is the magnitude of the contribution of a particular row element to the observed values of all units which it encompasses; the column effect is the magnitude of the contribution of a particular column element to the observed values of all units which it encompasses; and the interaction effect is the magnitude of the contribution due to the coming together of a particular row element with a particular column element.

The effects, the over-all value, and the errors of observation are assumed to be independent; their sum determines the various observed values. Since the mean of the errors of observation is assumed to be zero, and since the values within any one cell are assumed to vary only as a result of these measurement errors, the average value of X_{ijk} over a great number of replications within any one cell would be expected to be equal to the sum of its row, column, and interaction effects plus the over-all value. (In this paper

the over-all value in the case of the components of variance model is set equal to zero.)

For the most part, the above discussion applies only to the pure case in which replications are actually exact repetitions of each of the rc conditions. In many cases it is not feasible to have these exact replications within each cell because of effects of such factors as learning and motivation. But in using different (although comparable) units within the various cells, one introduces sampling errors in addition to the measurement errors. These sampling errors, however, can be made self-compensating, in the sense that they make an equal contribution (on the average) to all of the mean squares, by appropriate sampling and design. The examples in the following sections are illustrative of the ways in which sampling errors may be made self-compensating. If the sampling errors are handled in this way, the analysis is reducible to the general model established above.

As an illustration, let us take an experiment in which the row factor consists of individuals (the row elements being the specific individuals) and the column factor consists of a series of tests of visual acuity (the column elements being the specific tests). Let us assume an observed value or score of 30 for the second replication in the cell intersected by the third row and the first column (i.e., $X_{312} = 30$). This observation or score is made up of a number of components which sum to produce the specific value. We assume for the sake of exposition that the component contribution uniform for all observations in the third row is equal to 11 (i.e., the third row effect is equal to 11), that the component contribution of the visual acuity test in the first column for all of the observations which it encompasses is equal to 8 (i.e., the first column effect is equal to 8), that the component contribution of the unique interaction between the individual in the third row and the test in the first column for all of the observations in the ($r = 3, c = 1$) cell is equal to 3 [i.e., the ($r = 3, c = 1$) cell interaction effect is equal to 3], that the over-all value is equal to 6, and that the error involved in the second replication in the ($r = 3, c = 1$) cell is equal to 2. Thus, the assumption of the analysis of variance implies that the individual in the third row taking the test in the first column for the second replication will obtain the specified observed value or score of $11 + 8 + 3 + 6 + 2 = 30$.

The essential difference between the linear hypothesis model and the components of variance model is that the main effects of the former are fixed and constant whereas the main effects of the latter are random variables. All other differences result from the different mathematical treatments necessitated by this distinction. In order to have fixed and constant effects it is necessary that the elements of each factor be unique and not determined by random sampling; in order to have random effects the elements of the factors must be selected by simple random sampling from a larger population. Thus, if the entire population of elements is included in a particular factor,

its effects are fixed; if a random sampling of elements from some larger population is included in a particular factor, its effects are random. Examples of these kinds of effects will be presented later. The mixed model has one factor with fixed effects and the other factor with random effects.

When one makes a statistical test on the row effects he is testing an hypothesis of the type: Among the elements of the row factor (whether these be fixed or random) there is no variation in the magnitude of the contribution to the obtained observations. In other words: All the row effects are equal. Similarly for the column and interaction effects.

I. Linear Hypothesis Model

For purposes of exposition it will be convenient to divide the linear hypothesis model into three cases: (a) no a priori assumption as to interaction is made; (b) the a priori assumption is made that the interaction effects are equal, but no preliminary test of this assumption is desired; and (c) a preliminary test of the assumption of no interaction is made.

Case (a): Linear hypothesis model; no interaction assumption.

We assume

$$X_{ijk} = \mu_{ij.} + \mu_{i..} + \mu_{.j.} + \mu + \epsilon_{ijk}; \quad (1)$$

where

$\mu_{ij.}$ = fixed interaction effect, ($i = 1, \dots, r$), ($j = 1, \dots, c$);

$\mu_{i..}$ = fixed row effect, ($i = 1, \dots, r$);

$\mu_{.j.}$ = fixed column effect, ($j = 1, \dots, c$);

μ = over-all value (most commonly called the general mean).

That is, we assume that a particular observation is determined by the sum of the over-all mean, the effect of the row of which it is a member, the effect of the column of which it is a member, the effect of the row by column interaction for the cell of which it is a member, and an error term.

We further assume that the ϵ_{ijk} are independent random variables, normally distributed with mean equal to zero and variance equal to σ^2 (unknown).

We also have the assumptions

$$\sum_{i=1}^r \mu_{ij.} = 0, \quad \sum_{i=1}^r \mu_{i..} = 0, \quad \sum_{j=1}^c \mu_{ij.} = 0, \quad \sum_{j=1}^c \mu_{.j.} = 0. \quad (2)$$

These latter restrictions do not involve any loss of generality; if the effects which sum to make up X_{ijk} do not meet these assumptions, new values,

which meet these restrictions as well as the additive assumption, may be derived linearly from the original ones. That is, suppose we assume only that

$$X_{ijk} = \mu_{ij.} + \mu_{i..} + \mu_{.j.} + \mu + \epsilon_{ijk};$$

we may then derive

$$\mu' = \frac{1}{rc} \sum_{i=1}^r \sum_{j=1}^c u_{ij.} + \frac{1}{r} \sum_{i=1}^r \mu_{i..} + \frac{1}{c} \sum_{j=1}^c \mu_{.j.} + \mu, \quad (3)$$

$$\mu'_{i..} = \frac{1}{c} \sum_{j=1}^c \mu_{ij.} + \mu_{i..} - \frac{1}{rc} \sum_{i=1}^r \sum_{j=1}^c \mu_{ij.} - \frac{1}{r} \sum_{i=1}^r \mu_{i..}, \quad (4)$$

$$\mu'_{.j.} = \frac{1}{r} \sum_{i=1}^r \mu_{ij.} + \mu_{.j.} - \frac{1}{rc} \sum_{i=1}^r \sum_{j=1}^c \mu_{ij.} - \frac{1}{c} \sum_{j=1}^c \mu_{.j.}, \quad (5)$$

$$\mu'_{ij.} = \mu_{ij.} - \frac{1}{c} \sum_{j=1}^c \mu_{ij.} - \frac{1}{r} \sum_{i=1}^r \mu_{ij.} + \frac{1}{rc} \sum_{i=1}^r \sum_{j=1}^c \mu_{ij.}, \quad (6)$$

so that the derived values satisfy all of the assumptions and may be called the "effects."

(Note that we make no assumptions as to the existence or vanishing of the interaction effects for this case; i.e., the $\mu_{ij.}$ may be equal to 0 or to $K_{ij.}$, where $K_{ij.} \neq 0$, for all i, j .)

An example of this model would be an experiment in which c types of psychotherapy (perhaps directive versus nondirective or interpretive versus suggestive versus reflective) are employed by all r psychotherapists of a particular clinic in an attempt to have various subjects recall certain repressed material. Under each of the rc conditions there are m subjects; each subject is used under only one set of conditions (making the total number of subjects rcm). The recorded score in each case would be the time required for the particular therapist, with the particular therapeutic technique and with the particular subject, to get this subject to recall spontaneously certain (controlled) repressed material. Notice that we can generalize our results no further than to the therapists in this particular clinic since we considered the therapists not as a random sample from some larger population but rather as being fixed and distinct. The r therapists constitute our row factor population. (This distinction will be clearer after the discussion relative to the components of variance model is read.) If the experimenter had wanted to generalize his results to all psychotherapists in a particular area, while still utilizing the same c therapeutic techniques, he would have had to select a random sample of r therapists from the entire population of therapists in the area under consideration. It is acceptable to use all of the psychotherapists in one particular clinic in this latter case only if the assumption may be made that the therapists in the clinic represent a random sample of all the therapists in the area to which the results will be generalized. This case would constitute a mixed model.

The likelihood-ratio test (which is closely related to maximum likelihood estimation) is widely used for testing hypotheses in statistics since this test has many optimum properties. To test the hypothesis that there is no difference or variation among the row effects, or, what is equivalent, that there are no row effects (all $\mu_{i..} = 0$), the likelihood-ratio test leads to the ratio (see Table 1)

$$\frac{s_r^2}{s_w^2}, \quad (7)$$

which is distributed as F with $(r - 1)$ and $rc(m - 1)$ degrees of freedom under the null hypothesis. [For proof see (8), pp. 59-60.]

Case (b): Linear hypothesis model; assumption of no interaction, but without preliminary test.

As in case (a), we assume

$$X_{ijk} = \mu_{ij.} + \mu_{i..} + \mu_{.j.} + \mu + \epsilon_{ijk},$$

where ϵ_{ijk} are independent random variables, normally distributed with mean equal to zero and variance equal to σ^2 (unknown). And also

$$\sum_{i=1}^r \mu_{ij.} = 0, \quad \sum_{j=1}^c \mu_{ij.} = 0, \quad \sum_{i=1}^r \mu_{i..} = 0, \quad \sum_{j=1}^c \mu_{.j.} = 0.$$

This time, however, we make the additional assumption that there are no effects due to interaction; that is, $\mu_{ij.} = 0$ for all $i = 1, \dots, r$ and $j = 1, \dots, c$. (Note that this assumption makes the two assumptions above in regard to the summing of the interaction effects over i and j redundant.)

In this case, the likelihood-ratio theory leads to the following quotient for testing the hypothesis that there are no row effects (or all $\mu_{i..} = 0$):

$$\frac{s_r^2}{s_{i+w}^2}, \quad (8)$$

which is also distributed as F , but with $(r - 1)$ and $(r - 1)(c - 1) + rc(m - 1)$ degrees of freedom (if the null hypothesis is true). (For proof see 6, pp. 220-224).

Thus we see how the appropriate term (according to the theory of likelihood-ratio tests) to be used in testing the existence of the row effects (or column effects by similar reasoning) depends on the accepted assumptions. If an experimenter's data fit the assumptions behind the linear hypothesis analysis of variance model, and he makes no assumption as to the existence or non-existence of interaction, he uses (7); but if the experimenter can assume on the basis of some a priori reasoning that no interaction exists with data which fit the assumptions behind the linear hypothesis model, he uses (8).

Case (c): Linear hypothesis model; the use of a preliminary test.

Now the question arises as to the advisability and acceptability of the procedure of testing the significance of the interaction term by means of

$$\frac{S_i^2}{S_w^2}, \quad (9)$$

(which will be referred to hereafter as the "preliminary test") before determining whether to use (7) or (8) for the final test. Thus, if the interaction term is significant when tested by (9), the within cells mean square is used as the error term; if the interaction term is not significant when tested by (9), the sum of the interaction and within cells sums of squares divided by their combined degrees of freedom is used.

This is a compromise procedure which was originally derived on an intuitive basis by applied scientists in an attempt to utilize their experimental results and past knowledge to raise the power of their statistical test. [The power of a test is defined as one minus the probability of a type II error. See (10, pp. 246-248) for a good discussion of type I errors, type II errors, and power in the analysis of variance.] With this procedure the two tests (preliminary and final F) are not statistically independent, since they are both made on the same set of data. Thus, certain dangers are introduced.

This lack of independence and mathematical neatness has led mathematical statisticians to shy away from this area of application until very recently (and indeed some apparently still condemn this whole process of making the preliminary and final tests on the same set of data). Since 1944, Bancroft (2), Mosteller (11), Paull (12), and Bechhofer (3) have made important contributions to the problems involved in making preliminary tests of significance. The surface has as yet, however, barely been scratched.

In accord with the current literature the three possible procedures will be referred to as "never pool" (involving no assumption as to the existence or non-existence of interaction with no preliminary test), "always pool" (involving the assumption of no interaction with no preliminary test), and "sometimes pool" (where the error term in the final F -test depends upon the results obtained in a preliminary test of the significance of the interaction mean square).

Before presenting the formal summarization of the rules of procedure, certain symbols for degrees of freedom will be defined to facilitate the discussion. Accordingly, let

$$n_1 = (r - 1),$$

$$n_2 = (r - 1)(c - 1),$$

$$n_3 = rc(m - 1),$$

$$n_4 = (r - 1)(c - 1) + rc(m - 1);$$

and let $F(\alpha_x; n_i, n_j)$ refer to the value which is exceeded by F with probability α_x under the null hypothesis for the degrees of freedom n_i (numerator χ^2) and n_j (denominator χ^2); i.e.

$$\Pr \{F \geq F(\alpha_x; n_i, n_j)\} = \alpha_x. \quad (10)$$

(The subscript of α , that is, x , may be equal to 1, 2, or 3; the particular usage will be explained later.)

For testing the row effects in the two-factor, m replications case the statistical procedure may be summarized as follows:

"Never Pool"

Reject $\mu_{i..} = 0$, if

$$\frac{s_r^2}{s_w^2} \geq F(\alpha_2; n_1, n_3).$$

Accept $\mu_{i..} = 0$ otherwise.

"Always Pool"

Reject $\mu_{i..} = 0$, if

$$\frac{s_r^2}{s_{i+w}^2} \geq F(\alpha_3; n_1, n_4).$$

Accept $\mu_{i..} = 0$ otherwise.

"Sometimes Pool"

Reject $\mu_{i..} = 0$, if

$$\frac{s_i^2}{s_w^2} \geq F(\alpha_1; n_2, n_3)$$

and

$$\frac{s_r^2}{s_w^2} \geq F(\alpha_2; n_1, n_3); \quad (11)$$

or if

$$\frac{s_i^2}{s_w^2} < F(\alpha_1; n_2, n_3)$$

and

$$\frac{s_r^2}{s_{i+w}^2} \geq F(\alpha_3; n_1, n_4).$$

Accept $\mu_{i..} = 0$ otherwise.

Let us examine the advantages and disadvantages of each and the conditions under which each may be used.

The "always pool" procedure (where the interaction effects are in fact non-existent) provides a uniformly more powerful F -test than the "never pool" procedure for equivalent type I errors. [A uniformly most powerful test is one which is more powerful than all other possible tests (a test being defined by its critical region) regardless of the alternative to the null hypothesis which is assumed to be true.] If the "always pool" procedure is used and there actually are interaction effects, the denominator in the final F -test will tend to be too large, and the test will give too many non-significant results when in fact the null hypothesis is *not* true. Increase in interaction effects increases this distortion without limit, so that the research worker may be working at the, say, 1/500 per cent level of significance although he thinks he is working at the 5 per cent level. [See Table 2, p. 74 in Bechhofer (3) for an indication of how bad this disturbance gets under various conditions.]

The "sometimes pool" procedure is an attempt to avoid errors of this sort; the preliminary test is expected to advise against pooling when the interaction is large. The "sometimes pool" procedure cannot be expected to eliminate this source of error (or disturbance) entirely. But it is useful if it keeps the type I error of the final F -test close to the level at which the investigator thinks he is working. For equivalent type I errors this procedure also makes the power of the final F -test greater than the power of the final F -test under the "never pool" test.

It will be convenient for further exposition to introduce a term which summarizes the over-all magnitude of the interaction effects. Let this be

$$\lambda = m \sum_{i=1}^r \sum_{j=1}^c \mu_{ij}^2. \quad (12)$$

λ equals zero only when the interaction effects are all equal to zero; it gets proportionately larger as the μ_{ij} deviate from zero.

When λ is large, power and error characteristics make the use of the "sometimes pool" test theoretically unjustified (3). It is even more precarious to use the "always pool" test under these conditions. Thus, when an investigator has no a priori evidence to indicate a particular value for λ , he uses the "never pool" test, the routine use of the "sometimes pool" procedure being theoretically unacceptable.

In those cases in which the experimenter has definite a priori reasons for the belief that λ is equal, or at least close, to zero (that is, all μ_{ij} approximately equal to zero), and at the same time wants a certain amount of protection from an inaccurate assumption, the use of the "sometimes pool" procedure can be justified and is advantageous.

But more is involved than the mere caution that the use of the "sometimes pool" procedure requires definite a priori evidence indicating zero interaction. Since there is no preliminary test in the case of the "never pool" and "always pool" tests, their power is completely determined once the significance level of the final F -test is selected (for a given design, a specific value of λ , and an assumed-as-true alternative to the main effects' null hypothesis). When the "sometimes pool" procedure is used, however, the selection of a particular significance level for the final F -test merely limits the power of the whole test; it does not completely determine this power. The power (under the same conditions as above) is specified only when both the final and preliminary significance levels for the F -test are established.

Every combination of preliminary and final significance levels (for fixed degrees of freedom) within the general category of the "sometimes pool" procedure yields a different test. The "always pool" and "never pool" tests may simply be thought of as special (or extreme) cases of the "sometimes pool" procedure, with preliminary significance level equal to zero and one, respectively. Accordingly, as the preliminary significance level of a "sometimes

pool" procedure is decreased, it approaches an "always pool" test; as the preliminary significance level is increased, the "sometimes pool" procedure approaches a "never pool" test. Although the power of the entire test is greater with smaller preliminary significance levels, these smaller levels provide less protection from the disturbance resulting from an error in judgment as to interaction (particularly at the intermediate values of λ). Conversely, although there is more protection from the potential disturbance in total test significance level with larger preliminary significance levels, *there is less gain in power over the corresponding "never pool" test.*

[These relationships of power and total test type I error to the level of the preliminary test are not monotonic for all conditions. The best tests to be recommended in this paper have definite advantages in both power and error characteristics over many alternate tests. Nevertheless, the relationship indicated above does hold for wide and important (for protection purposes) ranges (a) in the magnitudes of the degrees of freedom, (b) in the selected final significance level, (c) in the value of λ , and (d) in the possible alternative to the main effects' null hypothesis.]

Bechhofer (3), for this model, and Paull (12), for the components of variance model, have worked out compromise tests which involve minimum danger of erroneous conclusions over the widest possible (for a uniform procedure) ranges in the values of interaction, the various degrees of freedom, and the possible alternatives to the main effects' null hypothesis. The procedure involves the free selection of the significance level of the final F -test; the preliminary significance level is established (by appropriate rules) so as to provide a test with the most desirable characteristics for the fixed final significance level.

What follows in this section is directed toward elaborating the statistical rules for establishing, for a fixed final significance level, that preliminary significance level for F which leads to the specific "sometimes pool" test with the most favorable power characteristics and minimum disturbance in significance levels for the linear hypothesis model.

Following Bechhofer (3), let

$$a = \left[\frac{(r-1)(c-1)}{rc(m-1)} \right] F(\alpha_1; n_2, n_3) \quad (13)$$

$$b = \left[\frac{(r-1)}{rc(m-1)} \right] F(\alpha_2; n_1, n_3) \quad (14)$$

$$c = \left[\frac{(r-1)}{(r-1)(c-1) + rc(m-1)} \right] F(\alpha_3; n_1, n_4). \quad (15)$$

For fixed degrees of freedom a is completely determined by α_1 , b by α_2 , and c by α_3 . α_1 is the level of significance to be used for the preliminary test. α_2 is the level of significance for the final F -test which the experimenter

would use if the preliminary test advises against pooling. α_3 is the level of significance for the final F -test, which the experimenter would use if the preliminary test recommends pooling. Notice that α_2 defines the "never pool" procedure when $\alpha_1 = 1$ and that α_3 defines the "always pool" procedure when $\alpha_1 = 0$. The above conditions define a "sometimes pool" test whenever $0 < \alpha_1 < 1$, which is the situation that interests us at the moment.

Within the category of "sometimes pool" tests we make three distinctions (3, p. 26):

$$\begin{array}{lll}
 \text{class } A \text{ tests} & \text{when} & c > \frac{b}{a+1}, \\
 \text{borderline tests} & \text{when} & c = \frac{b}{a+1}, \\
 \text{class } B \text{ tests} & \text{when} & c < \frac{b}{a+1}.
 \end{array} \tag{16}$$

The particular "sometimes pool" test is thus automatically determined once the significance levels (α_1 , α_2 , and α_3) are selected.

The foregoing exhaust all of the possible relationships which may exist between c and $b/(a+1)$. These values are under the control of the experimenter in the sense that he is free to choose the α -levels of significance; the latter uniquely (for fixed degrees of freedom) determine a , b , and c . The usual procedure employed by investigators is to choose the same level of significance for both the preliminary and the final F -tests; in this way the test is specified uniquely (and without their knowledge) for them. Let us take an example. Suppose an investigator chooses the 1 per cent level of significance for the preliminary F -test and also for the final F -test, regardless of the outcome of the preliminary test. Suppose also, in this example, that $r = 4$, $c = 5$, and $m = 3$. The F -value at the one per cent level for $n_2 = (r-1)(c-1) = 12$, and $n_3 = rc(m-1) = 40$ is 2.66; for $n_1 = (r-1) = 3$, and $n_3 = rc(m-1) = 40$, this level is 4.31; for $n_1 = (r-1) = 3$, and $n_4 = (r-1)(c-1) + rc(m-1) = 52$, this level is 4.18. This gives

$$a = \left[\frac{(3)(4)}{(4)(5)(2)} \right] (2.66) = .798,$$

$$b = \left[\frac{(3)}{(4)(5)(2)} \right] (4.31) = .323,$$

$$c = \left[\frac{(3)}{(3)(4) + (4)(5)(2)} \right] (4.18) = .241.$$

Since

$$.241 > \frac{.323}{.798 + 1} \quad \left(\text{that is, } c > \frac{b}{a + 1} \right),$$

this procedure amounted to a class *A* test. But, as Bechhofer (3, see particularly Tables 2, 3, 4, 5A, 5B) has shown, class *A* tests do not have the most favorable combination of power and error characteristics for this model. Thus, without definite knowledge or awareness, the experimenter selected a generally inferior test by employing this rather widely used procedure.

After a very thorough evaluation of the power and error characteristics of the various types of "sometimes pool" tests, Bechhofer concluded that the borderline test was the over-all best bet in terms of relative assurance of freedom from erroneous experimental conclusions. The borderline test does, however, introduce a slight distortion in the whole test type I error when $\lambda = 0$. That is, if α_2 is the type I error of the "never pool" test which the experimenter ordinarily would use, the borderline test defined by $b(\alpha_2)$ and $c(\alpha_2)$ has the property that its type I error will be larger than α_2 , when $\lambda = 0$.

To illustrate how much larger this type I error gets let us consider one of Bechhofer's examples (3, p. 81). For $\alpha_2 = \alpha_3 = 0.05$ and for $n_1 = 2$, $n_2 = 2$, and $n_3 = 6$, the maximum type I error of the borderline test (that is, when $\lambda = 0$) is 0.0653. This distortion is just about the worst that can be encountered in the two-factor, m replications design since the type I error decreases with either increasing λ or with increasing n_3 (regardless of λ) and approaches the limiting value of $\alpha_2 (= \alpha_3)$ very rapidly.

Thus, as Bechhofer concludes, "There is strong justification for the use of the borderline test under the circumstances specified. By tolerating a small increase in size [type I error] the experimenter can achieve a relatively large gain in power ... [when $\lambda = 0$]. He is protected against large ... (λ) ... since the power never will drop below the power of the conventional 'never pool' test he ordinarily would use." (3, p. 112). The absolute gain in power of the borderline test over the "never pool" test is a function of the degrees of freedom, and is greatest for smaller values of n_3 .

In addition to the advantage in gain of power of the borderline test, we saw in the preceding paragraph that its use brings freedom from any of the gross disturbances in type I error discussed in previous sections of this paper.

The recommended procedure, then, where the experimenter has strong a priori reasons for believing that the interaction effect is zero (or very close to zero), and where he wants to have some protection from the catastrophic effects of a completely erroneous assumption, is as follows:

1. Establish the level of significance (α) for the final F -test which would ordinarily be used for a "never pool" test, and set $\alpha_2 = \alpha_3 = \alpha$.

2. Determine b from

$$b = \left[\frac{(r-1)}{rc(m-1)} \right] F(\alpha_2; n_1, n_3), \quad \text{where} \quad \alpha_2 = \alpha.$$

3. Determine c from

$$c = \left[\frac{(r-1)}{(r-1)(c-1) + rc(m-1)} \right] F(\alpha_3; n_1, n_4), \quad \text{where} \quad \alpha_3 = \alpha.$$

4. Determine a from

$$a = \frac{b}{c} - 1. \quad (\text{The borderline test.})$$

5. Since a is determined, the F -value for the preliminary test, which gives the most effective "sometimes pool" test, may be found by

$$F(\alpha_1; n_2, n_3) = \left[\frac{rc(m-1)}{(r-1)(c-1)} \right] (a)$$

6. Now that the three F -values are established, we can proceed with the rule of procedure defined previously for the "sometimes pool" test, with the understanding that the type I error will be slightly larger than anticipated.

In the example given above, thus, instead of using the one per cent level of significance for both the preliminary and final F -tests one would proceed as follows (again $r = 4$, $c = 5$, $m = 3$):

1. The one per cent level of significance will be used for the final test regardless of outcome of the preliminary test.

$$2. \quad b = \left[\frac{3}{(4)(5)(2)} \right] (4.31) = .323$$

Where 4.31 is the F -value at the 1 per cent level of significance for $n_1 = 3$ and $n_3 = 40$.

$$3. \quad c = \left[\frac{3}{(3)(4) + (4)(5)(2)} \right] (4.18) = .241$$

Where 4.18 is the F -value at the 1% level of significance for $n_1 = 3$ and $n_4 = 52$.

$$4. \quad a = \frac{.323}{.241} - 1 = .340$$

$$5. \quad F(\alpha_1; 12, 40) = \frac{40}{12} (.340) = 1.133$$

6. Now we reject $\mu_{i..} = 0$ if

$$\frac{s_i^2}{s_w^2} \geq 1.133 \quad \text{and} \quad \frac{s_r^2}{s_w^2} \geq 4.31$$

or if

$$\frac{s_i^2}{s_w^2} < 1.133 \quad \text{and} \quad \frac{s_r^2}{s_{i+w}^2} \geq 4.18;$$

and accept $\mu_{i..} = 0$ otherwise.

II. Components of Variance Model

Again X_{ijk} is the observation in the i th row ($i = 1, \dots, r$), the j th column ($j = 1, \dots, c$), and for the k th replication ($k = 1, \dots, m$). In this model the row effects, the column effects, and the interaction effects are all assumed to be random variables. And

$$X_{ijk} = U_i + V_j + W_{ij} + Z_{ijk} \\ (i = 1, \dots, r), \quad (j = 1, \dots, c), \quad (k = 1, \dots, m). \quad (17)$$

(Notice that there is no symbol representing the over-all value above, since it is equal to zero for this model.)

U_i = the i th row effect,

V_j = the j th column effect,

W_{ij} = the i th row by j th column interaction effect,

Z_{ijk} = an error term.

[Roman letters represent the effects in this model; the Greek letter μ (in various forms) represents the effects in the linear hypothesis model. This notation is in accord with the practice of mathematical statisticians to use Greek letters for parameters (population values) and Roman letters for random variables. In the components of variance model the variances involved are the parameters.]

We further assume all U_i , V_j , W_{ij} , and Z_{ijk} are independent and normally distributed, with the following population values:

U_i	ξ_u	σ_u^2	
V_j	ξ_v	σ_v^2	
W_{ij}	ξ_w	σ_w^2	(18)
Z_{ijk}	ξ_z	σ_z^2	

$$(\xi = \xi_u + \xi_v + \xi_w + \xi_z)$$

If we convert these to the values

$$R_i = U_i - \xi_u, \quad (19)$$

$$S_j = V_j - \xi_v, \quad (20)$$

$$T_{ij} = W_{ij} - \xi_w, \quad (21)$$

$$Q_{ijk} = Z_{ijk} - \xi_z, \quad (22)$$

the new terms are independent and normally distributed with the population values

	Mean	Variance
R_i	0	$\sigma_r^2 = \sigma_u^2$
S_j	0	$\sigma_s^2 = \sigma_v^2$
T_{ij}	0	$\sigma_t^2 = \sigma_w^2$
Q_{ijk}	0	$\sigma_q^2 = \sigma_z^2$

(23)

$$X_{ijk} = \xi + R_i + S_j + T_{ij} + Q_{ijk} \quad (24)$$

$$\sigma_x^2 = \sigma_r^2 + \sigma_s^2 + \sigma_t^2 + \sigma_q^2 = \sigma_u^2 + \sigma_v^2 + \sigma_w^2 + \sigma_z^2. \quad (25)$$

As stated previously, the components of variance model differs from the linear hypothesis model in that the row and column effects are considered to be random variables, not fixed constants. That is, the individuals, tests, or situations which constitute the two factors in the components of variance model are randomly selected from two larger groups.

An example of this would be the following study to test the hypothesis that among all of the schools in a particular county there is a real difference in the average reading ability (within each school) of the third graders. For the study, r schools are selected at random from all of the schools in the county, c books are selected at random from all of the third-grade readers used in the county, and cm third-grade students are randomly chosen from each of the r schools. The students are each asked to read aloud 500 words from one of the readers; their average reading speeds are the recorded values. (The passages from the books are, incidentally, also chosen at random). There are, thus, m observations in each cell representing the reading scores of m different children from the school in the i th row, each reading the book in the j th column. It is assumed that the schools have what may be called "contribution to reading speed" values for third graders of U_1, U_2, \dots, U_r , which constitute a sample of r from a normal distribution of such values among all of the schools. It is also assumed that the c books have "ease of reading" contribution magnitudes of V_1, V_2, \dots, V_c ; the latter contributions constitute a sample of c from a normal distribution, drawn independently of the first values. If r other schools had been chosen, the values U_1, U_2, \dots, U_r would have been different, and if c other third-grade readers had been chosen the V_1, V_2, \dots, V_c would have been different. [We are not interested in any "basic reading capacity" which the children may have independently of their particular school training in this design since the random sampling and placement of these children enables this capacity to make an equal contribution (on the average) to all of the sums of squares.]

Thus, the results of this study may be generalized to the population as a whole (all the schools in the particular county in the case of the row effects—the test of the above hypothesis—and the entire population of third-grade readers used in the county in the case of the column effects).

Notice that we referred to the U_i , V_j , and W_{ij} as the row, column, and interaction effects, respectively. As a result of this, the testing of the row effects (for example) for significance in this model amounts to a test of an hypothesis of the sort: In the population as a whole, from which the sample was drawn, there is no difference in the row effects among the individual elements. In the above example this means we test the hypothesis that there is no difference in (or variation among) the “contribution to reading speed” values of the various schools (the row effects) and/or that there is no difference in (or variation among) the “ease of reading” contributing values of the books (the column effects) in the two populations. Thus, with this model the null hypothesis that the row effects do not vary in the population takes the form σ_u^2 (or σ_r^2) = 0. ($\sigma_r^2 = \sigma_s^2 = 0$ in the case of the columns).

If we had called the R_i , S_j , and T_{ij} the row, column, and interaction effects, respectively, the type of hypothesis we test and the preliminary assumption would be analogous to that of the linear hypothesis model. First, we would test an hypothesis such as that there are no row effects, rather than that there is no row effect difference or variation. This follows from the fact that the R_i (which have means equal to zero) are all identically zero and thus non-existent when $\sigma_r^2 = 0$, since their distribution is concentrated at the point zero. Also, our preliminary working equation for this model, showing the additive composition of the observed values, would have included a term representing an over-all value as in the linear hypothesis case. This term would be ξ , and, as μ previously, it would represent an over-all or general mean. Referring to the U_i , V_j , and W_{ij} as the effects is consistent with common practice and emphasizes the distinction between the effects of the linear hypothesis model and those of the components of variance model, as well as the treatment differences necessitated by this distinction.

As was the case with the linear hypothesis model, the choice of the proper term for testing the null hypothesis (in this case $\sigma_r^2 = 0$) depends on the assumptions made relative to the interaction effects. If the investigator makes no assumptions as to the equivalence of the interaction effects he tests the hypothesis $\sigma_r^2 = 0$ by

$$\frac{s_r^2}{s_i^2},$$

which is distributed as F only when $\sigma_r^2 = 0$, the numerator being too large otherwise. (For a delineation of the proof see 10, pp. 345-346.) This is the “never pool” procedure. Note that the interaction mean square is the proper error term for this model under these conditions; the within cells mean

square used as the error term in the linear hypothesis model is not the correct error term here.

If the investigator has ample reason to make the assumption that the interaction effects are identical (that is, $\sigma_i^2 = 0$), he may use the "always pool" procedure and test the hypothesis $\sigma_r^2 = 0$ by

$$\frac{s_r^2}{s_{i+w}^2}.$$

[For the essential features of the proof see (10), pp. 345-346.] This is the same test as used in the "always pool" procedure for the linear hypothesis model. As in the case of the other model, too, this procedure provides a uniformly more powerful test when the assumption is true.

Again there is motivation for the use of a preliminary test of significance by reason of doubt as to the validity of the assumption concerning the interaction effects. Contrary to the linear hypothesis model, the use of the "always pool" procedure, when there is in fact an interaction effect variation, results in the final F -test denominator tending to be too small, with the test giving too many significant results when the null hypothesis is true. With increase in interaction effect variations, this disturbance increases without limit as before, so that the experimenter may think he is operating at the 5 per cent level of significance while he may actually be operating at the, say, 37 per cent level.

The same preliminary test is used as for the preceding model, namely,

$$\frac{s_i^2}{s_w^2}.$$

The rule of procedure for this model may be summarized as follows:

"Never pool"

Reject $\sigma_r^2 = 0$, if

$$\frac{s_r^2}{s_i^2} \geq F(\alpha_2 ; n_1, n_2).$$

Accept $\sigma_r^2 = 0$ otherwise.

"Always pool"

Reject $\sigma_r^2 = 0$, if

$$\frac{s_r^2}{s_{i+w}^2} \geq F(\alpha_3 ; n_1, n_4).$$

Accept $\sigma_r^2 = 0$ otherwise.

"Sometimes Pool"

Reject $\sigma_r^2 = 0$, if

$$\frac{s_i^2}{s_w^2} \geq F(\alpha_1 ; n_2, n_3)$$

and

$$\frac{s_r^2}{s_i^2} \geq F(\alpha_2 ; n_1, n_2);$$

or if

$$\frac{s_i^2}{s_w^2} < F(\alpha_1 ; n_2, n_3)$$

and

$$\frac{s_r^2}{s_{i+w}^2} \geq F(\alpha_3 ; n_1, n_4).$$

Accept $\sigma_r^2 = 0$ otherwise.

Paull has found that a class *A* test has the most desirable properties for this mathematical model. This class *A* test, which is described below, is recommended "as one which tends to stabilize the disturbances at intermediate values of [the ratio of the expected value of the interaction mean square to the expected value of the within cells mean square] while still taking advantage of a considerable portion of the possible gain in power at values of [this ratio] near one" (12, p. 544). When this ratio, which is analogous to the λ of the linear hypothesis model, is large, there is little disturbance with the "sometimes pool" tests, since pooling is almost never advised. Paull recommends this procedure as the best compromise between the lower preliminary *F* significance levels (5 per cent, etc.), which do little to counteract the possible disturbance in the type I error of the final *F*-test, and the higher preliminary *F* significance levels (70 per cent, etc.) which provide little increase in power over the "never pool" procedure.

As a matter of fact, the effective or total test significance level when $\alpha_2 = \alpha_3 = 0.05$ and the preliminary *F*-test is made at the same level ($\alpha_1 = 0.05$) is considerably above 10 per cent for wide ranges in the value of the ratio of the expected value of the interaction mean square to the expected value of the within cells mean square.

The recommended class *A* procedure consists of pooling the interaction and within cells mean squares when their ratio is less than $2F(.50; n_2, n_3)$; that is, accept $\sigma_i^2 = 0$, if

$$\frac{s_i^2}{s_w^2} < 2F(.50; n_2, n_3),$$

and then carry on with the final *F*-test using s_r^2/s_i^2 if $\sigma_i^2 = 0$ is not accepted, and s_r^2/s_{i+w}^2 if $\sigma_i^2 = 0$ is accepted. [Fifty per cent points for the *F*-distribution may be found in the tables compiled by Merrington and Thompson (9).]

Let us illustrate this procedure with fictitious data like that used to show the linear hypothesis "sometimes pool" procedure (one per cent level of significance for the final test with $r = 4$, $c = 5$, and $m = 3$). The *F*-value for the 50 per cent level of significance with $n_2 = (r - 1)(c - 1) = 12$ and $n_3 = rc(m - 1) = 40$ is equal to .961; the *F*-value for the one per cent level of significance with $n_1 = (r - 1) = 3$ and $n_2 = (r - 1)(c - 1) = 12$ is 5.95; and the *F*-value for the one per cent level of significance with $n_1 = 3$ and $n_4 = 52$ is 4.18.

We will reject $\sigma_r^2 = 0$ if

$$\frac{s_i^2}{s_w^2} \geq 1.922 \quad \text{and} \quad \frac{s_r^2}{s_i^2} \geq 5.95;$$

or if

$$\frac{s_i^2}{s_w^2} < 1.922 \quad \text{and} \quad \frac{s_r^2}{s_{i+w}^2} \geq 4.18.$$

The constant multiplier 2 is arbitrary and a smaller value may be used where the experimenter is willing to sacrifice some power in the final F -test for increased assurance against extreme disturbance in significance level. A simpler rule which, according to Paull, may be used when the degrees of freedom of both rows and columns are greater than 6 is to pool the interaction and within cells mean squares when their ratio is less than 2. This is approximately equivalent to the above procedure, for large degrees of freedom, and does not necessitate reference to the F -table.

III. Mixed Model

Let us examine the case where the experimental data fit neither the assumptions of the linear hypothesis model nor those of the components of variance model exclusively, but fit the assumptions of a combination of the two. This is commonly called a mixed model.

We assume for the mixed model that the effects of one of the factors (say the columns) have been obtained by the random selection of elements representing that factor, while assuming that each element of the other factor (say the rows) has a constant effect which is typical for that element (i.e., no random sampling but rather fixed effects).

We assume that each observation is composed as follows:

$$X_{ijk} = \xi + \mu_i + S_i + T_{ij} + Q_{ijk}, \quad (27)$$

again ($i = 1, \dots, r$), ($j = 1, \dots, c$), ($k = 1, \dots, m$) and where ξ , S_i , T_{ij} , and Q_{ijk} are derived in the same way as in the components of variance model above. The μ_i are constants such that

$$\sum_{i=1}^r \mu_i = 0. \quad (28)$$

S_i , T_{ij} , Q_{ijk} are normal, independent, random variables with the population values

	Mean	Variance
S_i	0	σ_s^2
T_{ij}	0	σ_t^2
Q_{ijk}	0	σ_q^2

(29)

As in the case of the linear hypothesis model, we may wish to test the hypothesis of no row effects ($\mu_i = 0$, for $i = 1, \dots, r$); or, as in the case of the components of variance model, we may wish to test the hypothesis that the column effects are identical ($\sigma_s^2 = 0$). An example of this model was presented incidentally above as a variation of the linear hypothesis example.

There are two schools of thought in the literature as to the proper error

term for testing the hypothesis of no random effect variation ($\sigma_i^2 = 0$ in the case presented above) when $\sigma_i^2 \neq 0$. On the one hand are those represented by Anderson and Bancroft (1, p. 340) who assume fixed interaction effects for all of the observations encompassed by a given random main effect. That is, they assume that the entire population of interaction effects for each of the random elements is included in the sample since the entire population of elements intersected by the random elements is so included. In the notation of this paper this assumption implies (in the case where the column effects are random and the row effects are fixed)

$$\sum_{i=1}^r T_{ij} = 0.$$

Those represented by Mood (10, p. 348), on the other hand, assume that the interaction effects of the observations encompassed by both the random effects (columns above) and fixed effects (rows above) may be treated as random sampling variables exactly as in the case of the components of variance model.

The model advocated by Mood leads to an expected value of the random (main) effects mean square which includes the term $m\sigma_i^2$, while the model advocated by Anderson and Bancroft leads to a random (main) effects mean square whose expected value does not include $m\sigma_i^2$. Thus, the expected value of the random column effects mean square above is $\sigma_a^2 + mr\sigma_i^2 + m\sigma_i^2$ under the Mood assumption, and $\sigma_a^2 + mr\sigma_i^2$ under the Anderson and Bancroft assumption. As a consequence of this difference, the proper error term for testing the hypothesis of identical column elements (when $\sigma_i^2 \neq 0$) is the within cells mean square in the case of the Anderson and Bancroft model and the interaction mean square in the case of the Mood model.

If the position of Mood is accepted (and this implies that the interaction effect resulting from the coming together of a specific random element and a specific fixed element shows random error variation) the rule of procedure for this model for testing both the hypothesis of no row effects ($\mu_i = 0$) and the hypothesis of no column effect variation ($\sigma_i^2 = 0$) is identical to the rule for the components of variance model. The pooling procedure recommended by Paull (12) is applicable to the mixed model under Mood's assumption when the main effects to be tested include the random variable.

The rule of procedure if the position of Anderson and Bancroft is accepted (which is consistent with the general scheme presented early in this paper and usually more defensible) differs from that of the components of variance model in only one respect. The error term for testing the hypothesis of no column effect variation, when $\sigma_i^2 = 0$ is rejected, is s_w^2 instead of s_i^2 . The rule of procedure for testing the hypothesis of no row effect is identical to the components of variance model (except, of course, that an hypothesis of the sort $\mu_i = 0$ is accepted or rejected in the case of the mixed model).

Where one wishes to test the fixed main effects with the Mood model or where one wishes to test either of the main effects with the Anderson and Bancroft model, no specific recommendations for a preliminary significance level can be made at this time. The most satisfactory pooling procedure in terms of minimum disturbance or deviation in the significance level at which the experimenter thinks he is working and maximum power has, as yet, not been worked out under these conditions. All that has been said for the previous models regarding (a) the motivations for using each of the tests, and (b) the dangers and necessary cautions in using the "always pool" and "sometimes pool" procedures, applies equally to this model. It is as true for this model as it is for the other two that an investigator should not use either of these latter two procedures unless he has strong reason to believe that there are no interaction effects or interaction-effect differences, as the case may be.

REFERENCES

1. Anderson, R. L. and Bancroft, T. A. Statistical theory in research. New York: McGraw-Hill, 1952.
2. Bancroft, T. A. On biases in estimation due to the use of preliminary tests of significance. *Ann. math. Stat.*, 1944, 15, 190-204.
3. Bachhofer, R. E. The effect of preliminary tests of significance on the size and power of certain tests of univariate linear hypotheses. Unpublished doctor's dissertation, Columbia Univ., 1951.
4. Edwards, A. L. Experimental design in psychological research. New York: Rinehart, 1950.
5. Guilford, J. P. Fundamental statistics in psychology and education. New York: McGraw-Hill, 1950.
6. Johnson, P. O. Statistical methods in research. New York: Prentice Hall, 1949.
7. McNemar, Q. Psychological statistics. New York: Wiley, 1949.
8. Mann, H. B. Analysis and design of experiments. New York: Dover Publications, 1949.
9. Merrington, M. and Thompson, C. M. Tables of percentage points of the inverted beta (F) distribution. *Biometrika*, 1943, 33, 73-88.
10. Mood, A. M. Introduction to the theory of statistics. New York: McGraw-Hill, 1950.
11. Mosteller, F. On pooling data. *Jour. Am. stat. Assn.*, 1948, 43, 231-242.
12. Paull, A. E. On a preliminary test for pooling mean squares in the analysis of variance. *Ann. math. Stat.*, 1950, 21, 539-556.

Manuscript received 1/29/54

Revised manuscript received 4/27/54

A RATIONAL CURVE RELATING LENGTH OF REST PERIOD AND LENGTH OF SUBSEQUENT WORK PERIOD FOR AN ERGOGRAPHIC EXPERIMENT*

LEDYARD R. TUCKER

PRINCETON UNIVERSITY

AND

EDUCATIONAL TESTING SERVICE

A rational function is developed relating the length of a rest period and length of subsequent work period in an ergographic situation. Simple energistic postulates are used for a critical organ or neuromuscular structure whose failure to perform adequately results in a stoppage of the work period. Experimental results for two subjects using a finger ergograph indicate that the function yields the general trend of the data but that there seem to be some systematic deviations of the data from the present rational function. One parameter determined from the data represents rate of recovery from moderate fatigue. It is hoped that this development will aid in studies of motor functions as related to such other variables as age, motivation, and effects of drugs.

The idea for the present rational development occurred during a perusal of general literature on work decrement. A number of psychologists have used the ergograph in a variety of studies ranging from those concerned with personality characteristics to those dealing with work in industry. While considerable progress has been made by physiologists on characteristics of active muscles and nerves, there seems to have been only moderate success in application of these physiological developments to the problems encountered by psychologists in dealing with behavior of integrated, intact individuals. Indeed, there are a number of instances where psychologists claim that behavior such as exhibited with the ergograph cannot be accounted for on purely physiological and energistic grounds. The difficulty may be in finding how the various physiological details can be incorporated into descriptions of behavior of the complete individual. A second possibility is that psychologists have not considered sufficiently simple and limited behavioral situations to observe the physiological and energistic determiners of behavior. In the present case a few simple energy relations are postulated which only approximate the relations that might be determined on physiological grounds. These simple relations, however, permit development of a functional relation observable in the performance of an individual in a limited ergographic experiment. Psychologists

*This research was jointly supported in part by Princeton University and the Office of Naval Research under contract N6onr 270-20.

may find the present development of use in studying more complex situations.

After an individual has performed a constant, repetitive, motor task to such an extent that he no longer can continue, a rest period will result in the individual's being able to perform the task again for some work period before again being unable to continue. A graph relating length of rest period and

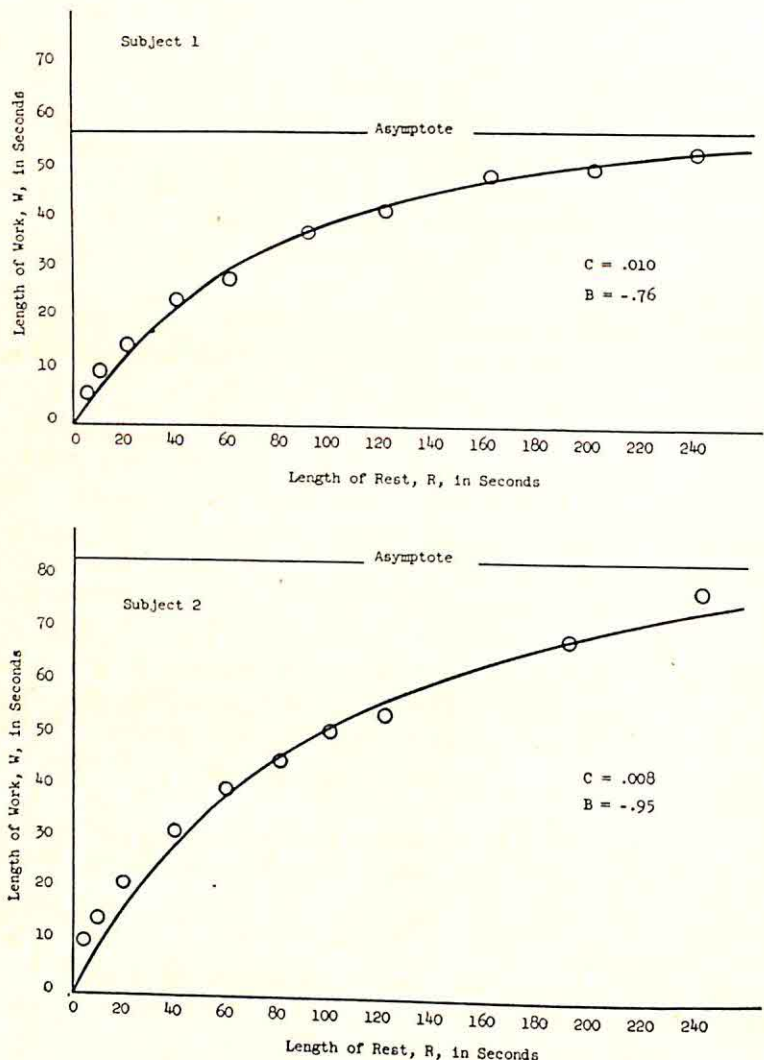


FIGURE 1
Rest-Work Curves for Two Subjects

length of subsequent work period will be of the form of those shown in Figure 1. A similar result was obtained empirically by Manzer (2). Short rest periods will be followed by short work periods, longer rest periods will be followed by longer work periods. As the rest periods are lengthened, the subsequent work

periods should also lengthen, but to a progressively lesser extent until some maximum length of work period is approached.

In developing a rational function, several assumptions are made concerning energy relations within the organ or neuromuscular structure whose failure to function adequately is responsible for the work stoppage. The organ might be the working muscle, or it might be one of the nervous elements responsible for excitation of the muscle. We will consider at this time only the critical organ or structure whose failure to function adequately results in failure of the individual to perform the task. It is assumed that fatigue of other organs or structures will have little effect on length of the work period so long as these organs do function. This is probably an oversimplification of the situation. Interaction between organs or structures probably does occur such that fatigue in one results in greater expenditure of energy by others in order for the individual to continue the task. This interaction is being ignored in the present development.

Consider an organ that is using energy at some constant rate during a working period. The supply of energy immediately available to the organ to be used in performing the task is being depleted. If this energy is being replenished at a slower rate than that at which it is being used, the supply of energy will be reduced. When the energy level falls to some critical point, the individual will be unable to continue the task and the work period will end.

During a rest period the energy supply of the organ will be replenished to an extent dependent on the length of the rest period and the rate at which energy is being made available to the organ by the rest of the individual's body. (For the present development, the nature of the physiological mechanism involved is not of immediate relevance.) At the end of such a rest period, the immediately available energy supply of the organ will again support performance of the task during a subsequent work period.

Consider the following postulates and definitions. Let:

E_t = energy immediately available to organ at time t ; (1)

E_m = energy immediately available to organ when it is in a completely rested state; (2)

a = rate of expenditure of energy during work period (postulated to be a constant); (3)

$C(E_m - E_t)$ = postulated rate at which body replaces energy to the organ; (4)

W = length of work period; and (5)

R = length of rest period. (6)

It is to be noted that postulate (3) forms a limit on the type of situation to which the present development is appropriate. The task should not be one for which the individual may work faster when more rested and then slow down when he becomes fatigued, nor should the task vary with the fatigue of the individual. The common type of ergographic series, where there may be long initial strokes followed by short strokes as the individual tires, is inappropriate

for the present development. In an ergograph situation the strokes should be of constant length and made at constant timing. Inability to make a stroke of standard length is to be interpreted as failure to perform the task. Thus, the individual is not driven to such fatigue that he cannot make a stroke of any length; he just cannot make one of standard length. Even in this case, this assumption of a constant rate of expenditure of energy is probably an approximation.

Postulate (4) involves the simple concept that energy replacement occurs at a rate proportional to the extent of deficiency below a maximum amount of energy available. This maximum amount of energy available is that which would be present in a completely rested organ. C is the constant of proportionality. $(E_m - E_t)$ is the extent of energy deficiency. This postulate is probably a gross approximation to a true function which could be determined from physiological considerations, but it should be usable for cruder developments and for cases involving a limited task, such as the flexion of a finger. This postulate would probably be inappropriate for more extensive tasks involving a large proportion of the body.

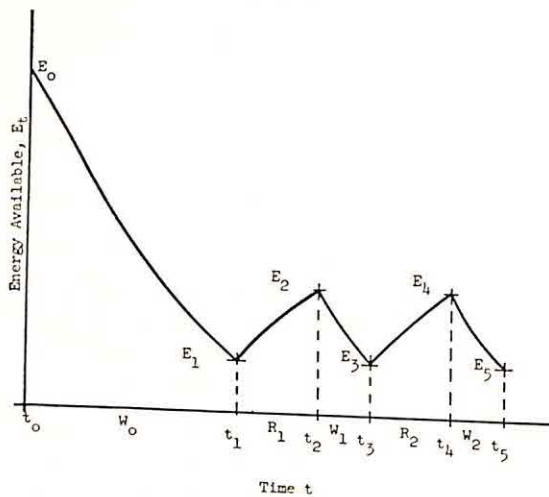


FIGURE 2
Relation Between Energy Available and Time During a Series of
Work and Rest Periods

Consider the curve of energy available versus time in Figure 2. At time t_0 the energy level is E_0 . As an initial work period progresses the energy level decreases along the curve until it reaches the level E_1 . This energy level E_1 is the critical value between continuation in performance of the task and discontinuation of performance. Whenever the energy level is below E_1 there is insufficient energy available for the organ to continue its part in the performance of the task. Whenever the energy level is above E_1 there is sufficient

energy for the organ to continue its activity. The time interval for this initial work period is W_0 .

A rest period of duration R_1 is now imposed and the energy level builds up to E_2 . During the following work period the energy level reduces to E_3 , a critical level between continued and non-continued performance of the task. Since the task has not been altered, we might postulate that

$$E_3 = E_1. \quad (7)$$

The duration of this work period is W_1 .

Consider that a second rest period of duration R_2 is imposed, which is followed by a work period of duration W_2 . The terminal energy level E is again the critical level between continued and non-continued performance of the task, and, therefore, is equal to E_3 and E_1 . Let

$$R_2 = R_1. \quad (8)$$

These two rest periods started with the identical energy levels E_3 and E_1 as postulated in (7); thus, if the energy restoration conditions are identical as postulated in (4), the energy levels at the end of these rest periods should be identical, i.e.,

$$E_2 = E_4. \quad (9)$$

It might be expected, then, that the two subsequent work periods would be identical also, i.e.,

$$W_2 = W_1. \quad (10)$$

This logic would lead to an expectation that a long sequence of rest-work periods with equal rest periods would have equal work periods. Yochelson (3) has reported data indicating such constancy in work periods following definite rest periods in long sequences of rest and work periods. Data gathered in the experimental try-out of this development also tended to support this contention. A finger ergograph working against spring tension with a fixed excursion to a block was used. The rate of contractions was set at one contraction per second. Preliminary trials revealed considerable initial practice effect, session to session, for a subject. During practice sessions some long sequences of rest and work periods with equal rest periods were tried out. The following results for the sixth practice session for one subject using a series of 60-second rest periods are typical. The lengths of the work periods, in seconds, were 54, 36, 30, 30, 31, 30, 28, 28, 28, 28, 33, 31, 30, 34, 28, 31, 39, 28, 28. The first one of 54 seconds in this series should not be counted. It corresponds to the initial work period W_0 before any of the fixed rest periods and might be expected to be long. The remaining work periods seem to vary within a fairly constant band with no apparent progressive decrement. Presumably this will hold only for a finite time and the experiment should not involve excessive sessions.

During the rest period R_1 the rate of change of energy with time can be obtained from postulate (4). Only energy replacement is considered to be active during the rest period.

$$\frac{dE_t}{dt} = C(E_m - E_t). \quad (11)$$

Integration yields

$$E_m - E_t = e^{(-Ct+f)}, \quad (12)$$

where f is a constant of integration. When the terminal times t_1 and t_2 and the corresponding energy levels E_1 and E_2 are substituted into (12), one obtains

$$\frac{E_m - E_1}{E_m - E_2} = \frac{e^{(-Ct_1+f)}}{e^{(-Ct_2+f)}} \quad (13)$$

$$= e^{(-Ct_1+f+Ct_2-f)} \quad (14)$$

$$= e^{C(t_2-t_1)}. \quad (15)$$

It is to be noted that the length of the rest period is

$$R = t_2 - t_1, \quad (16)$$

since t_2 and t_1 are the end and beginning times. The subscript to R is being dropped for convenience. Then (15) can be written

$$\frac{E_m - E_1}{E_m - E_2} = e^{CR}; \quad (17)$$

or, solving for E_2 ,

$$E_2 = E_m - (E_m - E_1)e^{-CR}. \quad (18)$$

Consider the subsequent work period. Energy is being used at a constant rate as per postulate (3) as well as being replenished as per postulate (4). Thus,

$$\frac{dE_t}{dt} = -a + C(E_m - E_t). \quad (19)$$

Integration yields

$$E_m - E_t - \frac{1}{C}a = e^{(-Ct+g)}, \quad (20)$$

where g is a constant of integration. Substitution of limiting times t_2 and t_3 and energy levels E_2 and E_3 and writing a ratio yields

$$\frac{E_m - E_2 - \frac{1}{C}a}{E_m - E_3 - \frac{1}{C}a} = \frac{e^{(-Ct_2 + a)}}{e^{(-Ct_3 + a)}} \quad (21)$$

$$= e^{C(t_3 - t_2)} \quad (22)$$

$$= e^{CW}, \quad (23)$$

where

$$W = t_3 - t_2. \quad (24)$$

Substituting E_1 for E_3 as per (7) and solving for E_2 yields

$$E_2 = E_m - \frac{1}{C}a - \left(E_m - E_1 - \frac{1}{C}a\right)e^{CW}. \quad (25)$$

In relating the work period and rest period, the two expressions for E_2 in (18) and (25) are equated to yield

$$E_m - (E_m - E_1)e^{-CR} = E_m - \frac{1}{C}a - \left(E_m - E_1 - \frac{1}{C}a\right)e^{CW}. \quad (26)$$

Subtracting E_1 from both sides of the equation,

$$E_m - E_1 - (E_m - E_1)e^{-CR} = E_m - E_1 - \frac{1}{C}a - \left(E_m - E_1 - \frac{1}{C}a\right)e^{CW}. \quad (27)$$

Or,

$$(E_m - E_1)(1 - e^{-CR}) = \left(E_m - E_1 - \frac{1}{C}a\right)(1 - e^{CW}), \quad (28)$$

$$\left(\frac{E_m - E_1}{E_m - E_1 - \frac{1}{C}a}\right)(1 - e^{-CR}) = (1 - e^{CW}). \quad (29)$$

Define

$$B = \left(\frac{E_m - E_1}{E_m - E_1 - \frac{1}{C}a}\right). \quad (30)$$

Then

$$B(1 - e^{-CR}) = (1 - e^{CW}). \quad (31)$$

Or, solving for e^{CW} ,

$$e^{CW} = 1 - B + Be^{-CR}. \quad (32)$$

Thus e^{C^W} is linearly related to e^{-C^R} with a slope of B and intercept of $1 - B$.

It is interesting to note that when the numerator and denominator of the right side of (30) are multiplied by C ,

$$B = \frac{C(E_m - E_1)}{-a + C(E_m - E_1)} \quad (33)$$

Thus, from (11) and (19)

$$B = \frac{\frac{dE_t}{dt} \text{ (for the rest period)}}{\frac{dE_t}{dt} \text{ (for the work period)}} \quad (34)$$

Another point of interest is that the relation given in (31) or (32) does not involve directly the amount of energy expended. Only measures of duration of rest and work periods need be determined. It is not necessary to observe the energy expended as is frequently attempted in ergographic experiments by computing the work performed by the muscle. (In case the muscle is not the critical organ responsible for the work stoppage, the work performed by the muscle would not be equal to the energy expenditure to be considered in case it were necessary to determine the constant a .) This fortunate feature is due to the restriction to a situation for which there is a constant rate of energy expenditure for the critical organ.

Three experimental sets of data were obtained. All used the finger ergograph previously described, which involved a spring load rather than weights. The excursion of the finger tip was limited by a block. The rate of finger contractions was set at one per second in all three cases. On each contraction, the limit block was to be touched. Failure to make a complete stroke ended each work period. In each experiment one subject was used for a number of sessions. Each session was composed of a "warm-up" period involving three work periods separated by 60-second rest periods. The first experimental rest period followed immediately the last warm-up work period. In the experimental session proper, a sequence of rest, work, rest, work, etc. periods were used. Instead of having a sequence of equal rest periods and thus determining one point on the rest-work curve at each session, each of the selected rest periods was used once at each session and the subsequent length of work period was determined. The order of rest periods was varied between sessions. Mean length of work periods following each length of rest period was determined for each subject.

Data for the initial subject are not presented here because he showed considerable practice effect from session to session. The subject performed about twice as much work in the fourth session as in the first session. The other two subjects received more extensive practice sessions and showed less increase in work performed during the experimental sessions. Results for the

preliminary subject were analyzed and the curve of (32) fits these data about as well as it does the data for the two subjects reported here.

Mean lengths of work periods following the chosen rest periods are given in Table 1.

TABLE 1
Experimental Results
(All times are in seconds.)

Subject 1		Subject 2	
Length of Rest	Length of Work*	Length of Rest	Length of Work**
5	5.8	5	10.0
10	9.9	10	14.7
20	14.9	20	21.5
40	23.5	40	31.8
60	27.8	60	39.7
90	37.5	80	45.2
120	41.5	100	51.0
160	48.1	120	54.0
200	50.0	180	67.0
240	52.9	240	77.8

* Mean length of work periods over 8 sessions.

** Mean length of work periods over 6 sessions.

The values of B and C for each subject were determined graphically. In cases where more precise determinations of these constants are desired, some more precise statistical method of curve fitting might be used. In the present case we were interested in obtaining only the proper order of magnitudes of B and C and felt that there was an advantage in the graphical method

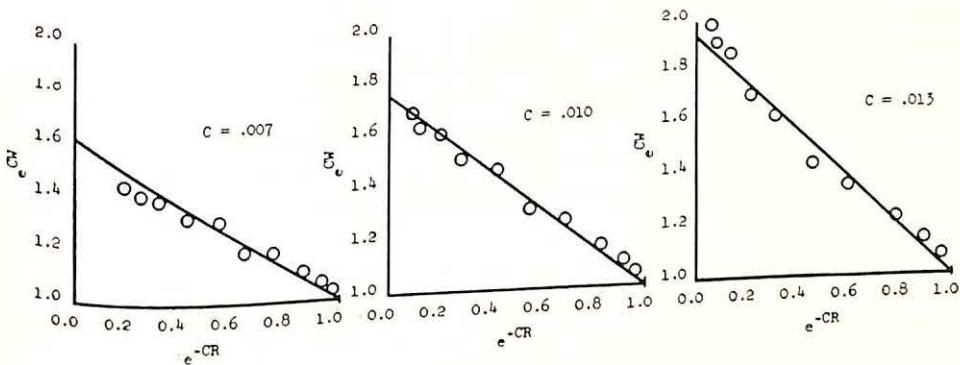


FIGURE 3
Rectification of Rest-Work Curve for Subject 1

in surveying the properties of the function and the data. A series of trial values of C were assumed. For each value of C , values of e^{CW} and e^{-CR} were obtained.

Figure 3 shows graphs for subject 1 for three values of C . Each point is determined by one rest period and the subsequent work period. From (32) it is expected that the points between e^{CW} and e^{-CR} would be linearly related for the proper value of C . Analysis of (32) also indicates that this line should pass through the point (1, 1). All three lines drawn in Figure 3 pass through this point. It is to be noted in Figure 3 that a low value of C yields a negative curvature and a high value of C yields a positive curvature. A C of .010 seemed to yield the best approximation to a straight line. A best-fitting line was drawn by eye with a slope of $-.76$, thus determining B . The line drawn on Figure 1 for subject 1 is the line for (32) with the values of B and C determined above.

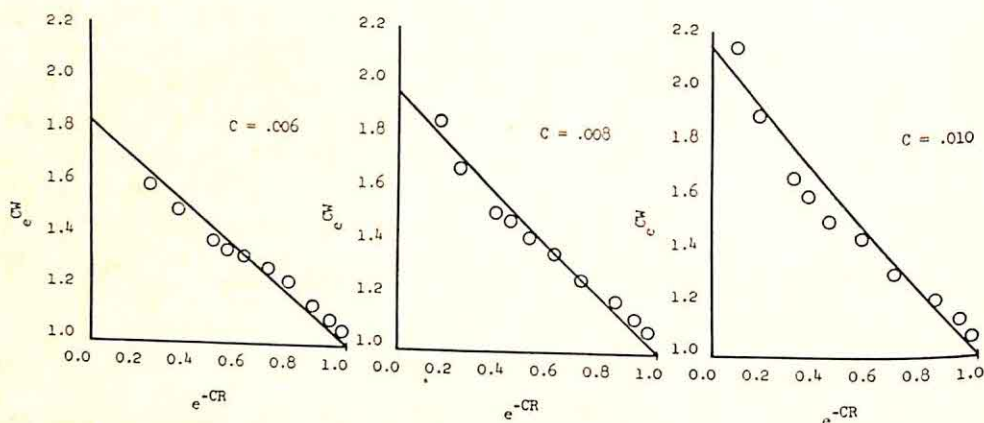


FIGURE 4
Rectification of Rest-Work Curve for Subject 2

Figure 4 shows graphs of e^{CW} and e^{-CR} for subject 2 and three trial values for C . While deviations from a straight line do not seem extreme, it is of interest that the points cannot be brought into more or less random fluctuations around a straight line by any choice of C . There seems to be a systematic wave about the straight line in each graph. This type of curve may result from the inadequacies of our approximations. A slight suggestion of this effect may be detected also for subject 1 in Figure 3. These results seem to be related to the results reported by Féré (1) and by Manzer (2), where performances after moderate rest periods were superior to performances after longer rest periods. While the consistency and seriousness of this lack of fit of the present function is a matter for further study, it is the author's judgment from the present data that (32) yields the general sweep of the observations. Some different set of postulates might yield a better fit to the data, or the systematic deviations may be considered as perturbations to be accounted for by further complexities in the mathematical model.

Returning to the fitting of (32) to the results for subject 2, values of .008 for C and $-.95$ for B seemed to give the best fit to the data. The corresponding curve is drawn in Figure 1.

An asymptote is indicated for each curve in Figure 1. This asymptote may be determined by setting R equal to infinity in (32.) Then e^{-CR} is zero and $e^{C'W}$ equals $(1 - B)$. In the experiment, the first work period exceeded this asymptote by some 10 to 20 per cent. This is a second indication of an inadequacy of our formulation which might be corrected by a more complex set of postulates. Another possibility is to interpret the present function to apply to the body state following the warm-up period in the experimental sessions.

Future work with the rational rest-work function can take any of several lines aside from development of a more adequate (and probably more complex) function. Individual differences in values of C and B for a fixed experiment might be correlated with other variables such as age. Effects of such conditions as ventilation, use of drugs, motivation, and response modality on C and B could be investigated. The experiment could be expanded to include systematic variation in load on the ergograph and timing of flexions, thus investigating other characteristics of the present function when a (rate of energy expenditure) and E_1 (critical energy level) are varied. It would be hoped that use of the rational function for the rest-work curve would help in obtaining greater precision in results for these various types of investigations.

REFERENCES

1. Féré, Charles. *Travail et plaisir*. Paris: Felix Alcan, 1904.
2. Manzer, C. W. An experimental investigation of rest pauses. *Arch. Psychol.*, 1927, 90, 1-84.
3. Yochelson, S. Effects of rest-pauses on work decrement. Unpublished Ph.D. dissertation, Yale University, 1930.

Manuscript received 2/2/54

Revised manuscript received 4/27/54



A MEASURE OF INTERRELATIONSHIP FOR OVERLAPPING GROUPS*

BEN J. WINER

PURDUE UNIVERSITY

A coefficient of interrelationship between overlapping groups that avoids indeterminacies inherent in the construction of fourfold tables for such purposes and, at the same time, is relatively insensitive to the absolute magnitude of marginal totals of fourfold tables, is developed. Under assumptions that are consistent with the objectives of organizational analysis, this coefficient is shown to be equivalent to a product-moment correlation coefficient. The advantages and limitations of this coefficient are pointed out. A numerical example giving computational procedures is presented.

Suppose an organization G of individuals is made up of k overlapping groups g_1, g_2, \dots, g_k . No restriction is placed upon the number of groups to which an individual may belong. Problems arise in analyzing the structure of organizations in which the equivalent of a correlation coefficient is needed. For example, one may seek means for simplifying the group structure of a complex organization. If one could construct the equivalent of an inter-correlation matrix, the factorial structure of this matrix might suggest means for restructuring the groups within the organization in such a way as to preserve many of the optimal conditions that may be present in the more complex structure and, at the same time, suggest ways of reducing the number of groups necessary to accomplish the same general mission.

As a starting point for this analysis, suppose one had the matrix of observations X , whose elements n_{ij} represent the number of individuals in G who belong to both g_i and g_j , i.e.,

$$X = \begin{bmatrix} n_{11} & n_{12} & \dots & n_{1k} \\ n_{21} & n_{22} & \dots & n_{2k} \\ . & . & \dots & . \\ n_{k1} & n_{k2} & \dots & n_{kk} \end{bmatrix}. \quad (1)$$

(Joint occurrence matrices of higher order will not be considered in the present development. Third-order matrices would have as elements n_{ijm} ,

*This measure was developed in connection with a study made by Dorothy C. Adkins (1). Her influence in the development was felt in many ways. The article was prepared while the author was employed at The University of North Carolina.

i.e., the number of individuals who are members of g_i , g_j , and g_m simultaneously.)

The number of individuals in each of the groups, n_{ii} , may show considerable variation. Up to a certain point, for purposes of analyzing organizational structure, the relative magnitudes of the n_{ii} are not particularly important. The coefficient of correlation sought, therefore, is one that is relatively insensitive to the magnitudes of the n_{ii} .

One of the simplest approaches to the problem would be to define the correlation between g_i and g_j by

$$r_{ij} = \frac{n_{ij}}{\sqrt{n_{ii}} \sqrt{n_{jj}}} \quad (2)$$

Although this definition of a correlation coefficient is equivalent to a product-moment correlation coefficient under rather general conditions, it has the disadvantage of being quite sensitive to the relative magnitudes of n_{ii} and n_{jj} . This disadvantage rules against its adoption.

An alternative approach might be to set up a fourfold table in order to compute phi coefficients, tetrachorics, or other types of coefficients of the same ilk. There are many possible fourfolds that can be constructed, depending upon what total of the marginal entries is considered appropriate. Adkins (1) humorously discusses the difficulties that arise when one starts to construct this type of fourfold table; in a sense, the minus-minus cell of this type of fourfold is indeterminate. One possible fourfold table might be

		g_j	
		-	+
g_i	+	$n_{ii} - n_{ij}$	n_{ij}
	-	$N - n_{ii} - n_{ij} + n_{ij}$	$n_{ij} - n_{ij}$
		$N - n_{ii}$	n_{ij}
			N

where N is the total number of individuals in G . If N is very much larger than $n_{ii} + n_{ij}$, most of the frequency would be concentrated in a single cell of the fourfold (the minus-minus cell). Phi coefficients computed from this type of fourfold are in general quite small in magnitude, have upper limits that are functions of the magnitude of the ratio n_{ii}/n_{ij} , and are quite sensitive to the magnitudes of n_{ii} and n_{ij} relative to N . For purposes of analyzing underlying group structure, a phi coefficient obtained from this fourfold table would be of highly limited value. A tetrachoric coefficient computed from this type of fourfold would not have so many limitations as does the phi coefficient; such coefficients are, however, also sensitive to small changes in n_{ii} and n_{ij} . In general, one is quickly led to the conclusion that a fourfold

table does not provide a satisfactory starting point for an index of overlapping group structure. [For a more detailed discussion of problems encountered in using fourfold tables in related areas see (2, 3, 4).]

An index which, in a real sense, is equivalent to a product-moment correlation coefficient can be derived from the observation matrix X by a series of assumptions that are consistent with the objectives of the subsequent organizational analysis. As a first step in this derivation, one sets up a matrix P having as elements

$$p_{ij} = n_{ij}/n_{ii}.$$

If D_n is a diagonal matrix having as diagonal elements n_{ii} , the matrix P is given by

$$P = D_n^{-1}X. \quad (3)$$

In general, P will *not* be a symmetric matrix. The elements of P represent the proportion of the members of g_i that also belong to g_j .

As a second step in the derivation, let D_p be a diagonal matrix having as diagonal elements the lengths of the row vectors of P , i.e., an element of D_p is given by $\sqrt{\sum_i p_{ii}^2}$. In order to normalize the rows of P , premultiply P by D_p^{-1} to obtain

$$F = D_p^{-1}P = D_p^{-1}D_n^{-1}X = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix}, \quad (4)$$

where a_i is a normalized row vector. The rows of F can be thought of as unit vectors in a k -dimensional vector space. Assuming that the basis vectors for this space are unit orthogonal vectors (justification of this assumption will be given presently), the elements of the matrix

$$R = FF' \quad (5)$$

represent the cosines of the angles between pairs of vectors. The correlation between g_i and g_j can be defined by

$$r_{ij} = \cos(a_i, a_j) = a_i \cdot a_j', \quad (6)$$

i.e., the scalar product of two row vectors having unit length.

In the language of factor analysis

$$F = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \cdot & \cdot & \cdots & \cdot \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{bmatrix} \quad (7)$$

TABLE 1

Matrix X = Number of Individuals in g_i
Who Are also in g_j

g	1	2	3	4	5	6
1	200	50	100	50	150	20
2	50	200	100	50	50	100
3	100	100	400	50	50	100
4	50	50	50	400	300	100
5	150	50	50	300	800	200
6	20	100	100	100	200	2000

TABLE 2

Matrix P = Proportion of Individuals
in g_i Who Are also in g_j

g	1	2	3	4	5	6	d_i^2
1	1.00	.25	.50	.25	.75	.10	1.9475
2	.25	1.00	.50	.25	.25	.50	1.6875
3	.25	.25	1.00	.12	.12	.25	1.2163
4	.12	.12	.12	1.00	.75	.25	1.6682
5	.19	.06	.06	.38	1.00	.25	1.2502
6	.01	.05	.05	.05	.10	1.00	1.0176

TABLE 3

Matrix F = Matrix of Unit Row Vectors
Giving Representation of g_i
in Terms of an Orthogonal Basis

g	1	2	3	4	5	6
1	.72	.18	.36	.18	.54	.07
2	.19	.77	.39	.19	.19	.39
3	.23	.23	.91	.11	.11	.23
4	.09	.09	.09	.78	.58	.19
5	.17	.05	.05	.34	.89	.22
6	.01	.05	.05	.05	.10	.99

TABLE 4

Matrix $R = F'F$

g	1	2	3	4	5	6
1	1.00	.58	.63	.58	.71	.17
2	.58	1.00	.71	.45	.41	.47
3	.63	.71	1.00	.33	.28	.30
4	.58	.45	.33	1.00	.85	.29
5	.71	.41	.28	.85	1.00	.33
6	.17	.47	.30	.29	.33	1.00

TABLE 5 (Short Cut)

$X'X$

	1	2	3	4	5	6
1	77,900	42,000	77,000	84,500	176,500	94,000
2	42,000	67,500	80,000	62,500	97,500	246,000
3	77,000	80,000	195,000	75,000	115,000	267,000
4	84,500	62,500	75,000	267,000	392,500	311,000
5	176,500	97,500	115,000	392,500	797,500	603,000
6	94,000	246,000	267,000	311,000	603,000	4,070,400
$\sqrt{\text{diagonal}}$	279.106	259.808	441.588	517.204	893.029	2017.523

TABLE 6 (Short Cut)

Matrix of Divisors of Elements of $X'X$

	1	2	3	4	5	6
1	77,900	72,514	123,249	144,355	249,250	563,103
2	72,514	67,500	114,728	134,374	232,046	524,169
3	123,249	114,728	195,000	228,391	394,351	890,914
4	144,355	134,374	228,391	267,500	461,878	1,043,471
5	249,250	232,046	394,351	461,878	797,500	1,801,707
6	563,103	524,169	890,914	1,043,471	1,801,707	4,070,400

can be interpreted as a matrix of factor loadings. The rows of this matrix can be considered to represent the projections of the groups on a set of orthogonal reference vectors. If g_i had no overlap with any of the other groups, the i th column of F would have unity in the i th row and zeroes elsewhere. Hence the orthogonal reference vectors, or factors, represent the location of ideal, non-overlapping groups in the k -space. This interpretation of the factors can be considered a possible justification for the choice of an orthogonal set of reference vectors rather than an oblique set. The effective dimension of the common factor space may be less than k ; indeed, it is the objective of studies in this area to find this reduced dimension.

Still another approach to the derivation of this proposed index of overlapping group structure suggests possible limitations implicit in it. Consider the i th row of F as representing the profile of the joining behavior of the average individual in g_i . Then the index of relationship derived here can be interpreted as a measure of profile similarity between two average individuals. (The deviations of each of the profiles are essentially measured from a common base line in terms of a metric relatively insensitive to n_{ii} .) For those groups which are homogeneous with respect to belonging behavior (i.e., high intra-class correlation), this average individual closely approximates all individuals in the group. Where the groups are not particularly homogeneous with respect to belonging behavior, this average profile (and the index of relationship based upon it) has somewhat limited value.

As a purely descriptive coefficient, the question of sampling distribution does not arise. The exact sampling distribution of the coefficient proposed raises a difficult problem in multivariate analysis. If the number of groups is large and the number of individuals within each group is also large, in spite of the fact that the coefficient can assume only positive values, it appears reasonable to assume that the sampling distribution of the multiple correlation coefficient provides an approximate sampling distribution for the index proposed here.

Numerical Example

An interesting application of the proposed index is given in (1). A smaller numerical application is presented here. Suppose an organization G consists of 5000 members. Further, suppose each member of G may belong to any, all, or none of six groups g_i ($i = 1, \dots, 6$). Let the number of individuals belonging to both g_i and g_j be given by the element in the i th row and j th column of the matrix shown in Table 1. The elements in the main diagonal of this matrix represent the number of individuals in each of the groups.

The matrix P (Table 2) is obtained from the matrix X by dividing each row of X by the corresponding entry in the main diagonal of X . The rows of P can be regarded as vectors representing g_i ; the squares of the lengths of

these vectors, d_{ii}^2 , are obtained by squaring and summing the entries in each row of P .

Normalized (unit) row vectors (Table 3) are obtained from P by multiplying the rows of P by $1/\sqrt{d_{ii}^2}$. The sums of the squares of the entries in each row of F should total unity (within rounding error). The matrix F can be considered as a matrix of factor loadings. The entries in row i of F represent the cosines of the angles between g_i and a set of hypothetically independent groups represented by an orthogonal set of basis vectors. From F one generates a correlation matrix in the same manner in which a correlation matrix is obtained from a set of orthogonal factor loadings, i.e., by post-multiplying F by its transpose.

A short-cut procedure is to compute the matrix XX' . If e_{ij} is the typical element in that matrix, then an element of the matrix R is given by

$$r_{ij} = \frac{e_{ij}}{\sqrt{e_{ii}} \sqrt{e_{jj}}}$$

In Table 5 the matrix XX' is computed; in Table 6 the typical element is $\sqrt{e_{ii}} \sqrt{e_{jj}}$. The matrix R is obtained by dividing each element of Table 5 by the corresponding element in Table 6.

REFERENCES

1. Adkins, D. C. The simple structure of the American Psychological Association. *Amer. Psychologist*, 1954, 9, 175-180.
2. Carroll, J. B. The effect of difficulty and chance success upon correlations between items or between tests. *Psychometrika*, 1945, 10, 1-19.
3. Wherry, R. J. and Gaylord, R. H. Factor pattern of test items and tests as a function of the correlation coefficient: content, difficulty, and constant error factors. *Psychometrika*, 1944, 9, 237-244.
4. Wherry, R. J. and Winer, B. J. A method for factoring large numbers of test items. *Psychometrika*, 1953, 18, 161-179.

Manuscript received 5/26/54

Revised manuscript received 7/5/54

AN EXTENSION OF ANDERSON'S SOLUTION FOR THE LATENT STRUCTURE EQUATIONS

W. A. GIBSON*

CENTER FOR ADVANCED STUDY IN THE BEHAVIORAL SCIENCES

Anderson's solution for the latent structure equations is summarized and then extended in two ways so as to involve all items simultaneously.

Some time ago Lazarsfeld and Dudman (4) achieved a solution, by means of determinantal equations, for Lazarsfeld's latent structure equations. Recently their solution was extended by Anderson (1) in such a way as to involve matrix manipulations only. Both of these solutions have the advantage over that of Green (2) of avoiding the need for estimating unknown elements in the manifest matrices—the elements with recurring subscripts. They have the disadvantage, however, of using much less of the empirical data than does Green's solution. The purpose of this note is to indicate two ways in which Anderson's solution can be extended so as to involve more of the empirical data and thus compare more favorably with Green's solution in that regard.

The latent structure equations have been developed elsewhere (3) and will merely be restated here in matrix form:

$$R = L'VL, \quad (1)$$

and

$$R_k = L'VD_kL, \quad (2)$$

where R is the sample joint proportions matrix bordered by the manifest marginals, R_k is the sample triple proportions matrix for item k , bordered by the joint proportions involving item k , L' contains the latent marginals for all items and has its top row filled with 1's, V is diagonal and contains the relative class sizes in its diagonal cells, and D_k is diagonal and contains the entries from row k of L' in its diagonal cells. All diagonal cells but the first in R and R_k and all cells in row and column k of M_k are empty and would have to be estimated if those cells were directly involved in the solution. The order of R and R_k is $n + 1$, n being the number of items involved, and the rank of all matrices in (1) and (2) is m , the number of latent classes needed to account for the manifest data.

*This article was written while the author was employed at The University of North Carolina.

Lazarsfeld (3, p. 389) has defined a basic determinant of R as a determinant formed from the rows and columns of R in such a way as to include the first diagonal element in R but no other diagonal element. Thus no basic determinant in R contains unknown elements. A basic determinant of R_k would be analogously defined and would contain no unknown elements provided row and column k of R_k were not involved. It is here convenient to speak of the *basic sub-matrices* of R and R_k as being the matrices of the basic determinants. For the present purpose the basic sub-matrices will always be dealt with in pairs—one from R and the corresponding one from R_k . Consequently the further restriction will be imposed that neither row nor column k of either R or R_k may be involved in a pair of basic sub-matrices. Finally, we shall be concerned only with basic sub-matrices of order and rank m .

Let P and P_k represent such a pair of basic sub-matrices. Then, by virtue of (1) and (2),

$$P = L'_1 V L_2, \quad (3)$$

and

$$P_k = L'_1 V D_k L_2, \quad (4)$$

where L'_1 is a square matrix made up of the first row of L' and of $m - 1$ other rows from L' , L_2 is a square matrix made up of the first column of L and of $m - 1$ other columns from L . From the restrictions that have just been stated, it follows that no item is represented both in L'_1 and L_2 , and that item k is represented neither in L'_1 nor in L_2 . Item k is, however, represented in D_k . Because of its role in the formation of R_k , Lazarsfeld has referred to item k as the *stratifier* (cf. 3, pp. 391–392).

Anderson's solution is simply to form the matrix,

$$A = P^{-1}P_k = L_2^{-1}V^{-1}L_1'^{-1}L'_1VD_kL_2 = L_2^{-1}D_kL_2, \quad (5)$$

and then obtain the characteristic roots and the right-sided characteristic vectors of A to get D_k and $L_2^{-1}K$, where K is an arbitrary diagonal matrix and remains, for the moment, unknown. Post-multiplying (5) through by $L_2^{-1}K$ shows that $L_2^{-1}K$ gives the right-sided characteristic vectors of A and that D_k contains the latent roots of A . Thus,

$$AL_2^{-1}K = L_2^{-1}D_kL_2L_2^{-1}K = L_2^{-1}D_kK = L_2^{-1}KD_k. \quad (6)$$

Post-multiplying (3) by $L_2^{-1}K$ gives

$$PL_2^{-1}K = L'_1VL_2L_2^{-1}K = L'_1VK. \quad (7)$$

Thus L'_1 becomes available except for multipliers on its columns. These multipliers turn out to be simply the entries in the first row of L'_1VK , since the first row of L'_1 must contain only 1's. L'_1 is thus obtained from the relationship,

$$L'_1 = (L'_1 VK)(VK)^{-1}. \quad (8)$$

Given L'_1 , the matrix product VL_2 is obtained from (3) as follows:

$$VL_2 = L'_1{}^{-1}P. \quad (9)$$

Both V and L_2 can now be obtained because the first column in VL_2 contains the diagonal elements of V .

In this form the solution by Anderson involves only $2m - 1$ of the items—the $m - 1$ items represented in L_1 , the $m - 1$ items in L_2 , and the stratifier k . There are two ways in which Anderson's solution can be extended to involve all of the items. No unknown elements will be introduced into the manifest matrices that are used. One way is to use a composite stratifier consisting of some combination of any of the items that are not represented in L_1 or L_2 . The other way is to augment the basic sub-matrices (hence also L'_1) by additional rows representing all items not involved in L_2 or in the stratifier.

The composite stratifier will be considered first. Let the subscript $kl--$ stand for a combination of any of the items that are not involved in L_1 or L_2 . Then the sum of the corresponding triple proportions matrices is given by

$$\begin{aligned} R_{kl--} &= R_k + R_l + \dots = L'VD_kL + L'VD_lL + \dots \\ &= L'V(D_k + D_l + \dots)L = L'VD_{kl--}L. \end{aligned} \quad (10)$$

By analogy with (10) the latent structure equation for a basic sub-matrix in R_{kl--} is

$$P_{kl--} = L'_1VD_{kl--}L_2. \quad (11)$$

P_{kl--} can be used in Anderson's solution in exactly the same way as is P_k . In the special case where the subscript $kl--$ refers to all of the items not involved in L'_1 and L_2 , it turns out that there is only one possible P_{kl--} .

A pre-publication reviewer has suggested that any weighted sum, and not just the simple sum, of R_k and D_k matrices could represent a composite stratifier.

Now consider the second way in which Anderson's solution can be extended. Let P and P_{kl--} be augmented by additional rows representing all of the items that are not represented in L_2 or in the (single or composite) stratifier. Thus P , P_{kl--} , and L'_1 cease to be square, but (3) and (11) still hold, and all other matrices in those equations remain square.

Now form the matrix

$$\begin{aligned} B &= (P'P)^{-1}P'P_{kl--} = (L'_2VL_1L'_1VL_2)^{-1}L'_2VL_1L'_1VD_{kl--}L_2 \\ &= L_2^{-1}V^{-1}(L_1L'_1)^{-1}V^{-1}L_2^{-1}L'_2VL_1L'_1VD_{kl--}L_2 \\ &= L_2^{-1}D_{kl--}L_2. \end{aligned} \quad (12)$$

The last step of (12) is identical with that of (5), except that the stratifier may here be composite. Thus the solution is the same as for Anderson from this point on, except that (9) is replaced by

$$VL_2 = (L_1L_1')^{-1}L_1P \quad (13)$$

because L_1' is no longer square.

It is perhaps worth mentioning that this extension of Anderson's solution can be shown to have two least-squares properties. The first is that the matrix B in (12) is such as to minimize the sum of squared discrepancies between the matrices $P_{k1..}$ and PB . The second is that the matrix VL_2 in (13) is such as to minimize the sum of squared discrepancies between the matrices P and $L_1'VL_2$.

A few remarks may be in order as to which items should be involved in each of the three matrices L_1' , L_2 , and $D_{k1..}$. Perhaps the best way to proceed is to locate that basic sub-matrix in R which seems from inspection to have the most clear-cut rank m . Attention might next be given to the make-up of the stratifier. The principal requirement here is dictated by the role of $D_{k1..}$ in (12). $D_{k1..}$ contains the characteristic roots of B , and if the right-sided characteristic vectors of B are to be unique, all diagonal elements in $D_{k1..}$ must be distinct and non-zero. At times it may be better to use only one item as a stratifier in order to insure this distinctness. At other times a composite may serve better than any single item in giving fairly even spread to the characteristic roots of B . In any event, some trial and error may be involved in the formation of an acceptable stratifier, for it will not always be possible to predict the necessary latent marginals with sufficient accuracy for this purpose. The solution by Green has this same problem (2, p. 158). Finally, all items not involved in the chosen basic sub-matrix of R nor in the trial stratifier can be thrown into the extra rows of P and $P_{k1..}$.

After an appropriate P and $P_{k1..}$ have been formed according to the requirements mentioned in the previous paragraph, the computing steps are as follows: (1) compute $P'P$ and get its inverse; (2) form the matrix $(P'P)^{-1}P'P_{k1..}$ and obtain its characteristic roots ($D_{k1..}$) and right-sided characteristic vectors ($L_2^{-1}K$); (3) compute the matrix product $PL_2^{-1}K$ and divide each of its columns through by the first entry in that column to obtain L_1' ; (4) compute L_1L_1' and get its inverse; (5) form the matrix $(L_1L_1')^{-1}L_1P$ and divide each of its rows through by the first entry in that row to obtain L_2 ; (6) form the diagonal matrix V from the divisors of step (5).

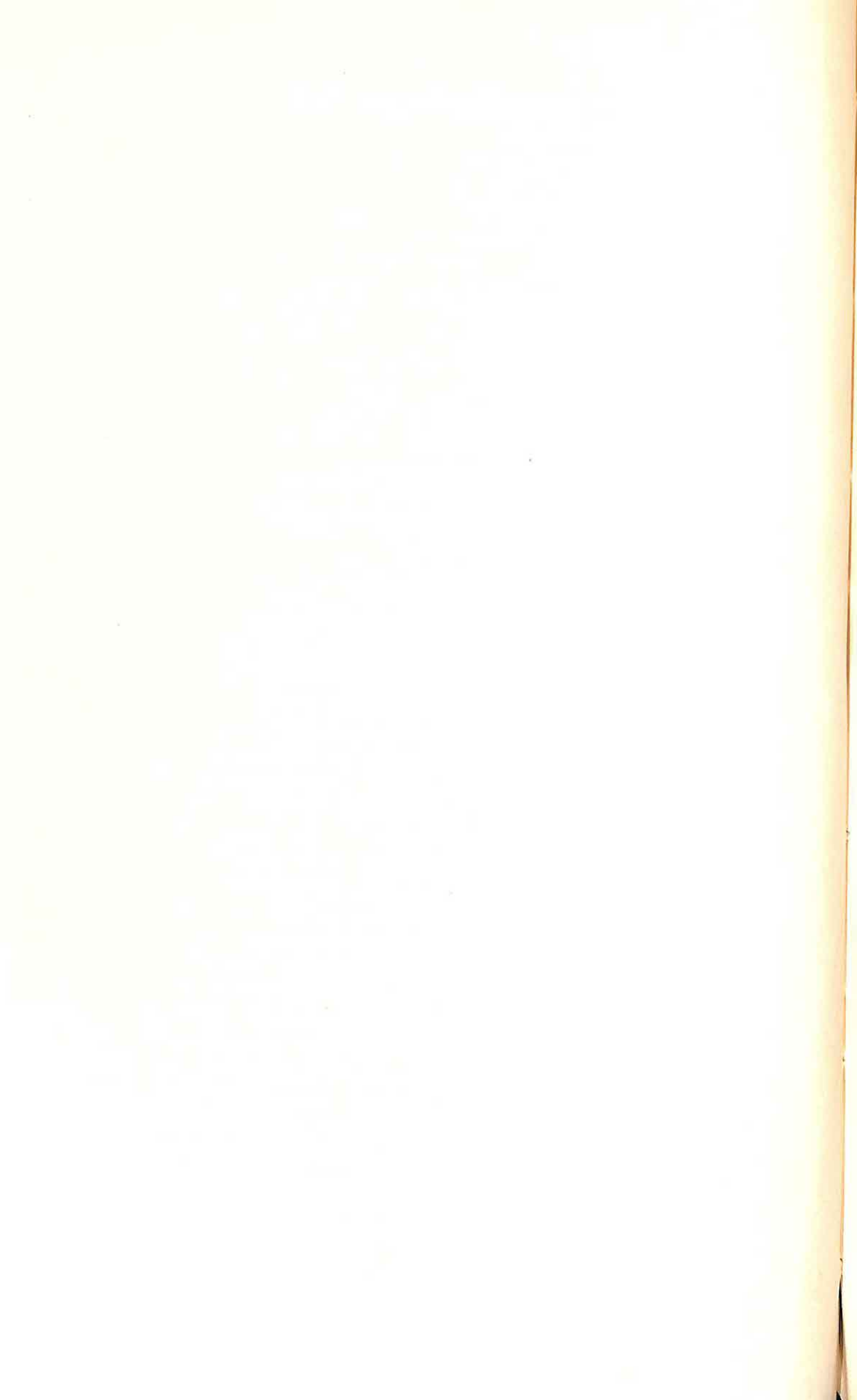
REFERENCES

1. Anderson, T. W. On estimation of parameters in latent structure analysis. *Psychometrika*, 1954, 19, 1-10.
2. Green, Bert F., Jr. A general solution for the latent class model of latent structure analysis. *Psychometrika*, 1951, 16, 151-166.

3. Lazarsfeld, Paul F. The logical and mathematical foundation of latent structure analysis. Chapter 10 in Stouffer, S. A., et al. Measurement and prediction. Princeton: Princeton University Press, 1950.
4. Lazarsfeld, Paul F., and Dudman, Jack. Paper No. 5 and Introduction to Paper No. 6 in Part II of Lazarsfeld, Paul F., et al. The use of mathematical models in the measurement of attitudes, Rand research memorandum no. 455, Santa Monica, The Rand Corporation, 1951 (mimeographed).

Manuscript received 5/4/54

Revised manuscript received 6/2/54



A FACTOR ANALYSIS OF MENTAL ABILITIES AND PERSONALITY TRAITS*

J. C. DENTON

PROCTER & GAMBLE

AND

CALVIN W. TAYLOR†

UNIVERSITY OF UTAH

The relationship between measures of verbal fluency and certain personality traits is examined by factor techniques. From a matrix of eight factor scores derived from mental tests plus five personality scores, six factors were obtained. An oblique solution lends limited support to the hypothesized relationship between the two domains.

In factorial studies of abilities, it has become general practice to include two or three "anchor" tests to measure each of the primary mental abilities that might be related to the experimental variables. In this way, one or more new factors may be isolated and interpreted with each successive well-planned study. The "anchor" tests which will probably best measure each of the several "established" factors can be identified fairly well.

This is not yet the case, however, for the area of temperament and personality, where there is much less agreement upon "anchor" variables. Using several different approaches, Cattell (1) has rather consistently found ten to twelve factors. Guilford has developed three inventories, STDCR, GAMIN, and I, which represent end products of his efforts to measure temperament factors. Although no serious effort has been made to compare the works of these authors, it seems that some of their factors may be quite similar while others apparently do not appear in both sets.

Few studies have straddled mental abilities and personality traits, even though the nature of any relationships found would be of considerable theoretical and practical importance. Thornton (9) found practically no overlap between tests of mental abilities and four questionnaire-type variables which measured a single factor called "Feeling of Adequacy." Other studies

*This study was supported by a grant from the Research Foundation of the University of Utah.

†Currently on leave of absence with the National Research Council.

likewise find little relationship (2). There is supporting evidence, however, for the hypothesis that fluent persons tend to be independent, extraverted, and unstable (7).

The present study is an effort to help define the relationships between mental abilities and personality traits, the latter being measured by a questionnaire. The major hypothesis was that there would be some relationship between measures of verbal fluency and extraversion or rathymia. Studies by Cattell and other British investigators (6) would tend to support this hypothesis. The relationship among the mental ability scores was also of interest since this involved, in a sense, a second-order factor study of eight cognitive factors. The personality score intercorrelations, which are admittedly distorted by experimental-dependence conditions in scoring (3), were of minor interest.

The Variables

Data on twenty-eight mental ability tests and on a personality inventory were collected by Taylor, the mental ability tests furnishing the basis for his study of fluency (8). For the present study, the fifteen tests were selected which best measured Taylor's eight primary abilities. Scores on two tests measuring the same factor were combined with equal weights to obtain a single index. (In the case of the Perceptual Speed factor, only one test was used.) These eight factor indices were included with five scores from a personality inventory for this study.

The eight factor indices, for which the tests are described by Taylor, were as follows:

1. *Memory* (First Names, Word-Number)
2. *Perceptual Speed* (Identical Numbers)
3. *Reasoning* (Letter Series, Letter Grouping)
4. *Number* (Addition, Multiplication)
5. *Verbal Comprehension* (Same or Opposite, Completion)
6. *Word Fluency* (First and Last Letters, Suffixes)
7. *Verbal Versatility* (Similes, Letter Star)
8. *Ideational Fluency* (Topics, Theme)

The remaining five personality variables from Guilford's "Inventory of the Factors STDCR" were:

9. *Social Introversion*
10. *Thinking Introversion*
11. *Depression*
12. *Cycloid Tendency*
13. *Rhathymia*

Procedures and Results

The data were obtained on 170 high-school seniors in Washington, D.C.

The score distributions on the eight factor scores were normalized. The matrix of correlation coefficients for the 13 variables was analyzed by Thurstone's group centroid method. Six factors were extracted and an oblique rotational solution obtained.

TABLE 1
Intercorrelations (above diagonal) and Residuals (below diagonal)

Var. 1	2	3	4	5	6	7	8	9	10	11	12	13	
1		04	35	21	22	19	14	17	-07	10	-10	-08	-04
2	-05		34	34	24	23	26	27	-20	07	01	05	22
3	01	06		33	51	36	40	33	-12	-12	-19	-10	14
4	02	02	-03		18	30	21	22	-16	-10	-11	-07	22
5	00	03	05	-03		36	40	38	07	01	-18	-18	-05
6	-01	-04	-02	03	02		32	32	-07	05	-01	07	07
7	00	-02	01	03	-01	-02		52	-16	11	07	16	27
8	02	-01	-01	03	-04	-01	-03		-17	16	-05	-01	10
9	-01	-03	-01	03	02	-01	-01	-01		06	45	23	-58
10	03	02	-01	-02	05	-01	-02	00	-01		53	48	-12
11	-01	-01	00	01	-03	-02	02	01	00	03		90	-08
12	00	-02	-01	00	00	03	-02	01	03	04	01		22
13	00	-01	-01	02	02	-04	02	-01	-08	-04	-02	-03	

TABLE 2
The Unrotated Factor Matrix

	I	II	III	IV	V	VI	h ²
1	38	-02	-10	03	38	-22	35
2	46	07	15	-05	-26	-15	34
3	70	-14	-01	-20	20	-01	59
4	48	-09	11	-18	-13	-33	41
5	63	-12	-27	00	01	22	53
6	55	07	-02	-04	-02	-01	31
7	62	16	15	09	-02	33	56
8	62	06	-01	28	-13	20	52
-9	20	-29	66	34	02	-16	70
10	06	63	-08	39	04	-10	57
11	-13	94	-11	-21	-07	03	96
12	-04	92	27	-26	05	12	100
13	21	02	76	-11	-07	05	65

The intercorrelations are presented in Table 1 along with the sixth-factor residuals (below the principal diagonal). The correlations among the personality scores were highest, as might be expected. In general, those among the mental abilities were next in size and those between mental abilities and personality traits were the lowest.

Table 2 presents the centroid matrix and Table 3 the factor matrix

TABLE 3
Final Rotated Matrix

	A	B	C	D	E	F	Variable
1	-.02	.47	.00	.03	-.06	.16	Memory
2	.40	-.10	.16	.06	.06	.06	Perceptual Speed
3	.11	.36	.33	-.06	.07	-.10	Reasoning
4	.46	.10	.00	-.03	-.01	-.03	Number
5	.04	.08	.57	-.12	-.14	-.04	Verbal Comprehension
6	.18	.11	.31	.07	.02	.05	Word Fluency
7	-.05	.00	.61	.09	.29	.12	Verbal Versatility
8	.06	-.10	.57	-.05	.06	.28	Ideational Fluency
-9	.00	.02	-.08	-.39	.51	.39	Social <u>Extraversion</u>
10	-.01	.02	.03	.50	-.03	.54	Thinking <u>Introversion</u>
11	.09	-.06	-.02	.94	.01	.00	Depression
12	-.05	.05	.01	.93	.42	-.01	Cycloid Tendency
13	.04	-.03	.04	.00	.71	.01	Rhathymia

TABLE 4
Final Transformation Matrix

	A	B	C	D	E	F
I	.27	.20	.57	-.01	.06	.13
II	-.24	-.18	.14	-.29	-.05	.90
III	-.68	.92	-.17	.10	.23	.11
IV	-.64	-.28	.78	-.08	.34	-.27
V	.04	.01	.02	.95	.12	.25
VI	-.07	-.03	-.15	-.04	.90	.13

TABLE 5
Reference Vector Cosines

	A	B	C	D	E
B	-.35				
C	-.24	-.27			
D	.09	.17	-.10		
E	-.40	.12	.12	.08	
F	-.08	.04	-.04	.00	.04

after rotation to simple structure. In these two tables and in all discussions hereafter, variable 9, Social Introversion, is treated as a reflected variable and is labeled - 9 and called Social Extraversion, for convenience. Table 4 presents the final transformation matrix and Table 5 the intercorrelations between the reference vectors.

The six factors in the rotated solution include three factors involving mental abilities and three mainly concerned with personality traits. None of

the personality variables had loadings on the mental ability factors, but two fluency variables did have loadings of almost .30 on each of two personality factors.

All variables with loadings greater than .25 are shown below for each factor. The interpretations are quite tentative, since the composite variables used are not like those customarily employed in factorial studies to define either primary or second-order factors.

Factor A

4. Number	.46
2. Perceptual Speed	.40

This is apparently a number, or perhaps a speed, factor. The appearance of Perceptual Speed (Identical Numbers test) is not surprising. In Taylor's original fluency study the same variable had a loading of .24 on the number factor. Tests designed to measure such abilities as Perceptual Speed or Carefulness which involve the manipulation of numbers frequently have significant projections on a number factor.

Factor B

1. Memory	.47
2. Reasoning	.36

Factor B is tentatively designated as a memory factor. Some of the major studies in the field have shown that tests of reasoning ability are related to memory (2, p. 148).

Factor C

7. Verbal Versatility	.61
8. Ideational Fluency	.57
5. Verbal Comprehension	.57
3. Reasoning	.33
6. Word Fluency	.31

This factor approaches a general factor of mental ability best represented by verbal tests, particularly by measures of fluency which involve the meaning of words.

Factor D

11. Depression	.94
12. Cycloid Tendency	.93
10. Thinking Introversion	.50
-9. Social Extraversion	-.39

This factor approaches a general (to this battery) measure of personality, each of the variables except Rhathymia having projections on it. It may be tentatively interpreted as Depression. Items dealing with "moodiness," "feelings easily hurt," "lost in thought," and "self-conscious" are typical of the depressive-type item contained in common in the *S*, *T*, *C*, and *D* scoring keys. Items such as these can account for much of the variance in this factor. The negative loading of variable 9, which was reflected prior to factoring, means that the unreflected variable, Social Introversion, is positively related to this factor. Factor *D* corresponds closely to Lovell's (4) factor which was called "Emotionality," or the opposite pole of Thurstone's (10) "Emotional Stability" factor.

Factor E

13. Rhathymia	.71
-9. Social <i>Extraversion</i>	.51
12. Cycloid Tendency	.42
7. Verbal Versatility	.29

Surgency is probably the best interpretation that can be given to this factor, in spite of the leading variable. Cattell has pointed out the similarity between Surgency and Rhathymia. This interpretation is supported by the positive loading of Social Extraversion and by the fact that it fits Studman's definition of a fluent person. In many ways it corresponds to the "Drive" factor found by Lovell. The loading of Verbal Versatility lends limited support to the original hypothesis. The correlation of .27 between Rhathymia and Verbal Versatility was larger than any other correlation cutting across the cognitive and personality domains. The other two types of fluency, Ideational Fluency and Word Fluency, showed no relationship with this "Surgency" factor.

Factor F

10. Thinking Introversion	.54
-9. Social <i>Extraversion</i>	.39
8. Ideational Fluency	.28

This rather ambiguous factor is not strongly determined and is difficult to interpret. Abstracting from Guilford's original definitions, this factor may represent "meditative thinking, philosophizing, and analyzing one's self" in addition to "entering into social contact, not shy" plus "fluent expression of ideas." Such a trait configuration seems somewhat unlikely. With present crude insights, it is difficult to sense what might be in common among these personality and mental ability scores. Again, it was a fluency score involving the meaning of words which showed a slight sign of bridging the gap into the personality domain.

Discussion

Since the mental ability variables analyzed here are all composites except one, the factor analysis results are similar in one sense to the second-order analysis reported by Rimoldi (5). Relationships in such studies generally seem magnified. This provides another reason for considering the interpretations of the factors as tentative.

The hypothesis that there is a relationship between fluency scores and certain personality characteristics is supported to a limited degree. The evidence relevant to this hypothesis is as follows: the fluency measure, Verbal Versatility, had a projection of .29 on the Surgency factor (*E*) and correlated .27 with Rhathymia. Ideational Fluency had a projection of .28 on an ambiguous personality factor (*F*); Word Fluency had zero loadings on the three personality factors, and all personality scores had zero loadings on the three mental ability factors. The results for the remaining mental ability variables are consistent with those of other investigations, in which many zero and a few low relationships between the mental ability and personality areas are reported. Improvement in test construction in both domains and further analyses may lead to higher correlations and also to greater insight into the bases of any relationships that appear.

REFERENCES

1. Cattell, R. B. *Personality: A systematic, theoretical and factual study*. New York: McGraw-Hill, 1950.
2. French, John W. The description of aptitude and achievement tests in terms of rotated factors. *Psychometr. Monogr.* No. 5, Chicago: The University of Chicago Press, 1951.
3. Guilford, J. P. When not to factor analyze. *Psychol. Bull.*, 1952, 49, 26-37.
4. Lovell, C. A study of the factor structure of thirteen personality variables. *Educ. psychol. Meas.*, 1945, 5, 335-350.
5. Rimoldi, H. J. A. The central intellectual factor. *Psychometrika*, 1951, 16, 75-101.
6. Rogers, C. A. A factorial study of verbal fluency and related dimensions of personality. *Amer. Psychologist*, 1952, 7, 290. (Abstract).
7. Studman, L. Grace. Studies in experimental psychiatry. V. 'w' and 'f' factors in relation to traits of personality. *J. ment. Sci.*, 1935, 81, 107-137.
8. Taylor, C. W. A factorial study of fluency in writing. *Psychometrika*, 1947, 12, 239-262.
9. Thornton, G. R. A factor analysis of tests designed to measure persistence. *Psychol. Monogr.*, 1939, 51, No. 229.
10. Thurstone, L. L. The dimensions of temperament. *Psychometrika*, 1951, 16, 11-20.

Manuscript received 1/18/54

Revised manuscript received 4/12/54

A TABULAR METHOD OF OBTAINING TETRACHORIC r WITH MEDIAN-CUT VARIABLES.

GEORGE SCHLAGER WELSH

THE UNIVERSITY OF NORTH CAROLINA

A method is presented that enables the immediate determination of tetrachoric r from a table if the proportion in the plus-plus cell for median-cut variables is known.

There is an ever-increasing use of factor analysis, cluster analysis, and related techniques in psychological research. Since numerous coefficients of correlation are required for the matrices, many investigators have employed tetrachoric r 's and have utilized various short-cut methods for obtaining these coefficients. This seems to be especially prevalent in preliminary investigations where the greater exactitude of more time-consuming methods of determining correlation is not feasible.

In many cases continuous distributions are dichotomized; it is often possible to make the cuts at the medians. The writer was able to divide at the median 24 of the 26 variables employed in a recent problem. To facilitate the determination of tetrachoric r a table was prepared so that the coefficient could be determined immediately if the proportion in the plus-plus cell were known (Table 1).

The table was prepared by using the computing diagrams of Chesire, Saffir, and Thurstone (1). To use these diagrams data are arranged in a fourfold table as follows:

	+	—	
+	c		b
—			
	a		

From the diagram for $a = .50$ the value of c corresponding to a particular value of r_{tet} was determined by noting where the r_{tet} curves from .10 through .95 cut the vertical line for $b = .50$. The proportions for 1.00 and .00 are, of course, known. These twelve points then described a curve with r_{tet} from .00 to 1.00 on the ordinate and proportions from .50 to .25 on the abscissa. These points were located on a large (26 by 30 inch) sheet of graph paper

TABLE 1

Three-Place Tetrachoric r Corresponding to Proportion in Plus-Plus
Cell for Median-Cut Variables

	Proportions									
	000	001	002	003	004	005	006	007	008	009
490	992	993	994	994	995	996	997	998	999	9995
480	984	985	986	987	988	989	990	990	991	992
470	972	974	975	976	978	979	980	981	982	983
460	958	960	961	962	964	965	966	968	969	971
450	943	944	946	947	948	950	951	953	955	956
440	924	926	928	930	932	933	935	937	939	941
430	904	906	908	910	912	914	916	918	920	922
420	878	881	884	887	890	893	895	898	900	902
410	847	850	854	857	860	863	866	869	872	875
400	814	818	821	824	827	831	834	837	840	844
390	777	781	785	789	793	797	800	804	807	811
380	734	739	744	748	753	757	761	765	769	773
370	688	693	697	702	707	711	716	721	725	730
360	639	644	649	654	659	664	669	674	678	683
350	589	594	599	604	609	614	619	624	629	634
340	536	542	547	552	558	563	568	573	578	584
330	482	487	493	498	504	509	515	520	525	531
320	426	431	437	443	448	454	460	465	471	476
310	368	374	380	385	391	397	403	409	414	420
300	309	315	321	327	333	339	345	351	356	362
290	249	255	261	267	273	279	285	291	297	303
280	187	194	200	206	212	218	224	230	237	243
270	125	132	138	144	150	157	163	169	175	181
260	063	069	075	082	088	094	100	106	113	119
250	000	006	013	019	025	031	038	044	050	057

Decimal points properly preceding each entry have been eliminated.

and a smooth curve drawn. From the graph the 250 three-place r_{tet} 's for the corresponding proportions were determined.

To insure accuracy a check was made by computing from formula the proportions for various r_{tet} 's lying between .10 and .80. When both variables are cut at the medians the usual formula for r_{tet} to five terms can be reduced and rearranged to solve for the proportion in the plus-plus cell as:

$$\text{proportion} = .15915504 \left(r_{tet} + \frac{r_{tet}^3}{6} + \frac{3r_{tet}^5}{40} \right) + .250.$$

These proportions and those obtained graphically agreed to two places with only rounding errors in the third place. Values above $r_{tet} = .80$ could not be checked by means of the shortened formula but this section of the curve was redrawn on a larger scale and the values checked.

Table 1 is used in the following way:

(1) when both variables are cut at the medians and arranged in a

fourfold table, determine the proportion to three places falling in the plus-plus cell;

(2) this proportion is found in the marginal entries of the table;

(3) entering the table read off the $r_{t,t}$ to three places from the body of the table (see example A).

(4) if the proportion in the plus-plus cell is less than .250, the $r_{t,t}$ will be negative. In this case use the proportion in the plus-minus cell (or .500 minus the plus-plus proportion) and place a minus sign before the obtained $r_{t,t}$ (see example B).

Examples:

A.

	+	-	
+	67	33	100
-	33	67	100
	100	100	200

$$\frac{67}{200} = .335$$

At the intersection of row 330 and column 005 read off $r_{t,t} = .509$

B.

	+	-	
+	40	110	150
-	110	40	150
	150	150	300

$$\frac{40}{300} = .133, \frac{110}{300} = .367$$

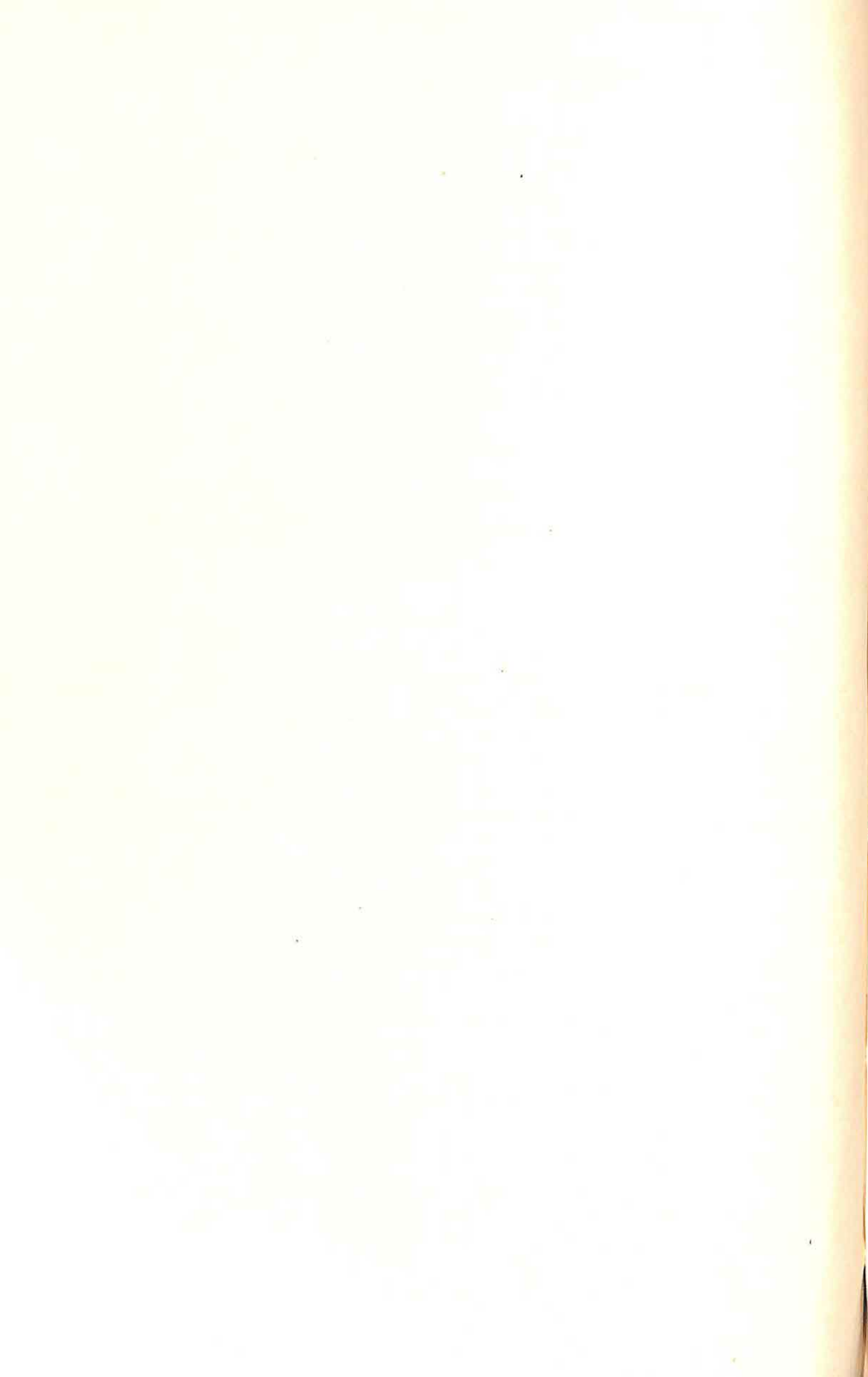
At the intersection of row 360 and column 007 read off $r_{t,t} = -.674$

REFERENCE

1. Chesire, L., Saffir, M. and Thurstone, L. L. Computing diagrams for the tetrachoric correlation coefficient. Chicago: Univ. Chicago Bookstore, 1933.

Manuscript received 2/4/54

Revised manuscript received 3/27/54



AN IBM METHOD FOR COMPUTING INTRASERIAL CORRELATIONS*

M. CARR PAYNE, JR.

AND

LEONARD STAUGAS

UNIVERSITY OF ILLINOIS

A method for computing intraserial correlations using a 602-A Calculating Punch, an 077 Collator, a 513 Gang Punch, and a 403 Tabulator is described. An example of the use of the procedure and an estimate of the time needed with each machine are given. This procedure is compared with another method, which makes use of a more powerful IBM machine.

Introduction

In a recent article, Grant (3) described an experimental approach to behavior as a time series which new developments and adaptations of quantitative techniques have made possible. In the fields of psychophysics and motor performance, the new techniques (1, 2, 6) typically have dealt with binary data, e.g., did a subject see a light flash or not. A more generally useful technique of time series analysis is one which uses continuous data. One such technique obtains correlations by calculating Pearson product-moment correlations within the series (intraserial correlations). An automatic method for computing these correlations with typical IBM equipment is described below.

Computing intraserial correlations requires the pairing of measures which were separated by any stated number of measures in the original series. Lag x (or τ_x) is the case in which an event is displaced x events from the one with which it is paired in obtaining a serial correlation (or autocorrelation). To find the correlation coefficient for lag 1 over a series of N measures, it is necessary to pair measure 1 with measure 2, 2 with 3, 3 with 4, \dots , $N - 1$ with N . The n for computing the correlation is equal to N minus the lag number. If the correlations are to be obtained with IBM machines of the order of the 602-A Calculating Punch, the data must be so organized that each score appears on the same punched card as the score with which it is to be paired in the correlation computation. The following procedure was worked out using the 602-A

*This research was supported in part by the United States Air Force under Contract No. AF 33(038)-25726, monitored by the Air Force Personnel and Training Research Center. Permission is granted for reproduction, translation, publication, use and disposal in whole and in part by or for the United States Government.

Calculating Punch, the 077 Collator, the 513 Gang Punch, and the 403 Tabulator. [A very interesting approach to this same problem, using a more powerful IBM calculator is described in Schipper and Gruenberger (5).]

Outline of the Procedure

In general terms, the procedure is as follows: The data are punched in conventional form across the card, several cards usually being needed to contain the measures for one series. Punching begins in a column which depends on the number of lags to be computed and the number of digits in each measure. (See the boxed-in entries on the original cards in Fig. 1). In addition to the measures themselves, each card contains appropriate identification of the data. A certain number of blank cards depending on the location of the first punched column, and the number of digits per measure are now inserted behind every original data card (see "inserted cards" in Fig. 1). Measures are punched from the original card into these blank cards using the interspersed master-card gang-punching principle. As this gang-punching is carried out, each measure is offset to the left by one measure on each succeeding card. Thus, if each score contains two digits, a score in columns 23 and 24 on the original card appears in columns 21 and 22 on the second card, in columns 19 and 20 of the third card, etc. This principle, which has been used by various workers, was described several years ago by Hartley (4). When the original cards and the inserted cards have been passed through the Gang Punch once, the first columns through the deck, depending on the number of digits per measure, contain all the measures in the series. In the two digit case these are contained in columns 1 and 2 through the deck. All cards which do not contain punches in these first columns are discarded.

The next step is to construct a set of "answer" cards, each of which is eventually to contain the computations relating to one correlation. An answer card is inserted at the end of each deck representing a series of measures. The whole deck is then passed through the 602-A Calculating Punch. The 602-A is wired to accumulate the N , $\sum X$, $\sum Y$, $\sum X^2$, $\sum Y^2$, and $\sum XY$ for the lag 1 correlation and to punch them into the answer card. To obtain similar information for lag 2, three steps are followed: (a) the whole file is put through the Collator, which in one operation removes the old answer card, removes the terminal data card on which the needed Y measure is missing, and inserts a new answer card, (b) the 602-A is rewired to accumulate the needed sums for lag 2, and (c) the file is passed through the 602-A, and the new answer card is punched. This procedure is repeated for as many lags as are to be examined in the series of measures.

When all the correlation answer cards have been assembled, they are put through the 602-A twice more, using two panels which compute the square of the Pearson product-moment coefficient, using the raw score formula. In the process, the squared correlations are punched into available unused columns

on the answer cards. The answer cards are then merged with a table of cards containing all possible values of r , and using the interspersed master-card gang-punching principle again, the proper r is found and punched directly on the answer cards. The answer cards are then ordered appropriately and a listing of the correlations is prepared on the Tabulator.

An Example

As an example of the use of this procedure we may cite the intraserial correlations computed for lags 1 through 10 for some brightness matching data. These data were available as two-digit scale readings. They were punched, 20 readings per card, beginning in column 23. Suitable identification was punched in columns 63-67. Different series of the data varied in the number of readings, but for a series of 120 readings, for example, it was necessary to punch six cards. The first card of each series was identified with "X" punches for interspersed master-card gang-punching. The Collator was used to insert 19 blank cards behind each original card except the last one of the series which was followed by 29.

Columns 3 through 62 of the punch brushes of the Gang Punch were wired, in order, into columns 1 through 60 of the punch magnets so as to accomplish off-set gang-punching from each card into the next. The columns containing identification were wired to gang-punch normally and the "PX" hubs wired for interspersed master-card gang-punching. The Gang Punch read the values on each card and punched them into positions two columns to the left on the succeeding card. By the time the first nineteen inserted cards had been punched, the data were displaced to the left far enough so that columns 1 through 22 on the next card, the second original card, were punched with values which serially preceded those in columns 23 through 62 on that card (as in Fig. 1). This off-set punching was continued until the last two responses of the data were punched in columns 1, 2, 3, and 4 of the last card in the deck.

After all the cards had been passed through the Gang Punch, the first 11 of each series (the original punched card and the first ten inserted cards) were removed and discarded as they did not contain punches in columns 1 and 2. At this point, each series deck was as follows: the first card (inserted card no. 11 in Fig. 1) contained the first through the 20th reading in the first 40 columns plus appropriate identification. The second card contained readings 2 through 20 in the first 38 columns. The last card contained the last two readings in the first four columns. It was now possible to obtain lag 1 correlation by working with columns 1 and 2 for the X value and columns 3 and 4 for the Y value, through the deck of cards. Measures 1 and 2 were in these columns on the first card, measures 2 and 3 on the second card, etc.

The 602-A Calculating Punch calculated and punched the N , sums, sums of squares, and sums of cross products for each correlation into an

appropriately identified answer card that was inserted at the end of each series. After all the correlation components had been obtained for a particular lag, the file of cards was put through the Collator to remove the old answer cards, insert answer cards for the next lag, and remove those score cards which were no longer appropriate. These inappropriate score cards were those which did not contain punches in all the columns being considered for the new Y values. This calculating and collating process was repeated for each of ten lags. To check on the efficiency of the Collator, it was found to be worthwhile to examine visually and to count all rejected score cards.

At this point, each answer card contained the components necessary for computing the correlation coefficient for a particular lag. These computations were carried out as described above, and the resulting r 's and components were tabulated.

Calculation Times

The time needed on each machine to calculate 280 correlations with N 's ranging from 119 to 110 (120 original responses correlated for 10 lags) in the study used in this example is summarized below.

Key Punch	.5 hour
Calculator	26.5 hours
Collator	3.0 hours
Sorter	1.0 hour
Tabulator	.5 hour
Gang Punch	3.0 hours
<hr/>	
Total time	34.5 hours

Discussion

This computing procedure provides no machine check on results obtained at any stage. In the work cited above, suspicious coefficients were recalculated through the entire procedure and also as a check, a few non-suspicious coefficients were randomly selected and calculated on a desk calculator. None of these checked coefficients or components was found to be in error. The procedure has the advantage that all of the components of each correlation coefficient are punched into one "answer card," which makes it easy to use these values in other calculations where they may be needed.

For data where each measure is known very precisely and contains a large number of significant digits, the procedure outlined by Schipper and Gruenberger (5) using a more powerful calculator is probably more desirable than the present one. However, in most psychological research the number of significant digits in each measure is small. In this case the difference in time per correlation between the two procedures does not warrant the use of the more high-powered calculator. The present procedure, for reasons of ready availability of the equipment needed, simplicity, and ease of understanding, is probably the more satisfactory one for most psychological research.

REFERENCES

1. Abelson, R. P. Spectral analysis and the study of individual differences in the performance of routine, repetitive tasks. Princeton: Educ. Test. Serv., 1953.
2. Flynn, J. P. Lack of randomness in sequences of auditory differential threshold data. *Amer. Psychologist*, 1948, 3, 254. Abstract.
3. Grant, D. A. The discrimination of sequences in stimulus events and the transmission of information. *Amer. Psychologist*, 1954, 9, 62-68.
4. Hartley, H. O. The application of some commercial calculating machines to certain statistical calculations. *Supp. J. roy. statist. Soc.*, 1946, 8, 154-183.
5. Schipper, L. M. and Gruenberger, F. A method of calculation of serial correlation coefficients utilizing the IBM Card-Programmed Electronic Calculator. *Res. Bull.* 53-10, 6564th Research and Development Group, HRRC, Air Research and Development Command, Lackland Air Force Base, Texas, May, 1953.
6. Verplanck, W. S., Collier, G. H., and Cotton, J. W. Nonindependence of successive responses in measurements of the visual threshold. *J. exp. Psychol.*, 1952, 42, 273-282.

Manuscript received 12/31/53

Revised manuscript received 4/12/54

ESTIMATION AND TESTS OF SIGNIFICANCE IN FACTOR ANALYSIS

C. RADHAKRISHNA RAO

VISITING RESEARCH PROFESSOR, UNIVERSITY OF ILLINOIS

A distinction is drawn between the method of principal components developed by Hotelling and the common factor analysis discussed in psychological literature both from the point of view of stochastic models involved and problems of statistical inference. The appropriate statistical techniques are briefly reviewed in the first case and detailed in the second. A new method of analysis called the canonical factor analysis, explaining the correlations between rather than the variances of the measurements, is developed. This analysis furnishes one out of a number of possible solutions to the maximum likelihood equations of Lawley. It admits an iterative procedure for estimating the factor loadings and also for constructing the likelihood criterion useful in testing a specified hypothesis on the number of factors and in determining a lower confidence limit to the number of factors.

1. *Introduction*

Whatever may be the arguments for or against factor analysis as a tool in psychological research, the statistical problems it involves have been of considerable interest to the statistician mainly because of their complexity. Two important contributions on the statistical side are by Hotelling (8), who introduced the principal component analysis, and Lawley (11, 12), who provided a test criterion for judging the significance of factors in addition to working out the maximum-likelihood equations of estimation. These two authors were, however, considering two different problems, both of which seem to have important application. They are sometimes considered as two possible formulations of the same problem providing the same answer. In theory it helps to make a distinction between the two. The term principal component analysis (PCA) should be used for Hotelling's formulation of the problem and its solution; the term factor analysis should be used for the specialized formulation considered in psychological literature and for the various solutions offered (see also 10). Lawley was considering the latter problem under the assumption that the variables (test scores) are normally distributed.

Illustrations have appeared from time to time to show that PCA gives nearly the same relative magnitudes of factor loadings as any effective method of factor analysis. This is true only when what have been termed as communalities are very nearly equal for all the tests as shown in section 3.1 of this paper.

The PCA is sometimes modified (3, p. 114; 7) by the insertion of communalities in the diagonal of the correlation matrix. This method, called the principal factor analysis (PFA), seems to provide a valid approach to the problem of factor analysis; however, it carries with it the flavor of principal component analysis intended to explain the variations in the standardized scores. It will be seen that an alternative approach developed in section 3.2 explains most effectively the correlations between the test scores in a battery. This method may be called a canonical factor analysis (CFA). Formulas for estimation are detailed in section 4.

The tests of significance associated with component analysis and factor analysis also differ to some extent. In the former case interest chiefly lies in the magnitude of, or the relationship between, the latent roots of the hypothetical matrix of raw correlations or those corrected for attenuation. In factor analysis it is the decomposition of the correlation matrix as the sum of a diagonal matrix and a positive semi-definite matrix. The differences in nature of these tests are sometimes fundamental (section 2.2). The tests of component analysis are contained in Hotelling's paper (8), and an appropriate test for factor analysis is given by Lawley (11). An alternative form of Lawley's test yielding slightly more precise results is given in section 4.2. It is also shown (section 4.3) that the test criterion can be calculated during the process of estimation and used in obtaining a lower confidence limit to the number of factors.

Recently Bartlett (1, 2) proposed a test involving the latent roots of the correlation matrix intended to study "the correlation structure in relation to the variance of the measurements." The exact nature of the hypothesis for which Bartlett's test is applicable and the conditions under which it is valid are examined in section 2.2. It appears that this test does not provide a complete answer to either form of analysis under consideration.

For a full account of tests of significance in factor analysis developed up to 1952, the reader is referred to Burt (3).

The author of this article is not concerned here in examining which of the methods, component or factor analysis, is relevant in problems of psychological research or whether both methods provide rather similar numerical results (not identical in general) leading to the same psychological interpretation. The main emphasis is on the differences in statistical techniques needed in these two cases and a detailed examination of the methods for factor analysis.

2. Problems of Factor and Component Analyses

2.1 Factor Analysis

Factor analysis postulates an underlying structure of a set of measurements in terms of hypothetical variables (non-observable) depending on

what are called common and specific or individual factors. If x_1, \dots, x_p denote p different measurements on an individual, then x_i is written

$$x_i = z_i + s_i \quad (i = 1, \dots, p), \quad (2.1.1)$$

where z_i , the variables depending on common factors, and s_i , the variables depending on specific factors, satisfy the following conditions of zero covariance:

$$\text{cov}(z_i, s_i) = 0, \quad \text{cov}(z_i, s_j) = 0, \quad \text{cov}(s_i, s_j) = 0 \quad (i \neq j). \quad (2.1.2)$$

Sometimes, another independent variable representing unreliability in the measurement x_i is added to $(z_i + s_i)$, but for purposes of factor analysis based on unrepeatable test scores of individuals, this variable can be combined with s_i without loss of generality. If such repeated test scores are available, then a more comprehensive analysis of the common and specific factors is possible. This latter analysis is not considered here.

From the structural setup (2.1.1), (2.1.2) it follows

$$V(x_i) = V(z_i) + V(s_i)$$

$$\sigma_{ii} = \gamma_{ii} + \delta_{ii}.$$

$$\begin{aligned} \text{cov}(x_i, x_j) &= \text{cov}(z_i, z_j) + \text{cov}(z_i, s_j) + \text{cov}(s_i, z_j) + \text{cov}(s_i, s_j) \\ &= \text{cov}(z_i, z_j) \end{aligned}$$

$$\sigma_{ij} = \gamma_{ij} \quad (i \neq j).$$

If Σ , Γ , Δ denote the dispersion matrices of the vector variables x , z , s , then

$$\Sigma = \Gamma + \Delta,$$

where Δ is a diagonal matrix.

It is seen from the above analysis that any correlation between x_i, x_j is solely due to the correlation between z_i, z_j . What we can actually observe are the values of the variables x on a group of individuals but not z, s which are not operationally defined, but whose hypothetical existence is postulated. We thus obtain an estimate of the matrix Σ . The subject of factor analysis is mainly concerned with the estimation of the matrix Γ starting with an estimate of Σ . The object is not to find any matrix Γ satisfying the condition $\Sigma = \Gamma + \Delta$ but the one which has the least complexity leading to a parsimonious description of the relationships between the observable variables x . The complexity, when defined as the rank of the matrix Γ , has a special significance for the problems on which this technique is applied, as shown in the subsequent sections of this paper.

Some of the statistical problems of factor analysis are:

(a) to estimate the minimum rank of the dispersion matrix (variances and covariances) Γ of the variables z_1, \dots, z_p occurring in the structural equations (2.1.1, 2.1.2),

- (b) to test any hypothesis specifying the minimum rank of Γ ,
- (c) to estimate a basis of the common factor space (defined below),
- (d) to predict the value of any common factor from the observed set x_1, \dots, x_p for any individual.

The statement that the rank of Γ is $k < p$ implies that the variables z_1, \dots, z_p can be expressed as linear combinations of k independent variables only. To bring out the precise meaning of such a dependence let us consider the entire vector space of elements consisting of all linear combinations of the set z_1, \dots, z_p of variables introduced into the different tests of a battery with the restriction that any two variables differing by a constant represent the same element. We may call any element of this space a common factor variable or simply a factor variable unless otherwise specified. The vector product of any two elements f and g of this space is defined by $\text{cov}(f, g)$ and the square of the norm of f by variance of f , $V(f)$.

Vectors f_1, f_2, \dots, f_r of this space are said to be independent if no linear combination $a_1 f_1 + a_2 f_2 + \dots + a_r f_r$ (all $a_i \neq 0$ simultaneously) has zero norm. A vector space is said to be finite dimensional if all its elements can be expressed as linear combinations of a finite number of elements. In such a case there is a minimal number of such elements called the rank of the space. A set of such elements necessarily independent (to be minimal) is called a basis. A basis of a vector space is not, however, unique but its rank is.

We can always choose a basis such that its elements are orthogonal (zero vector product), implying that the chosen factor variables are uncorrelated. A convenience provided by such a choice is that a basis can be simply represented by a set of correlations between the measurements and factors. If Z_1, \dots, Z_k is an orthogonal basis then each z_i can be expressed in terms of Z_j .

$$z_i = a_{i1}Z_1 + \dots + a_{ik}Z_k. \quad (2.1.3)$$

The covariance of x_i with Z_j is a_{ij} , the coefficient of Z_j in the representation (2.1.3). This may be regarded as a correlation coefficient once x_i and Z_j are properly standardized. A basic set of factors or equations (2.1.3) can be represented by the matrix of correlations

$$\begin{bmatrix} a_{11} & \dots & a_{1k} \\ \vdots & & \vdots \\ a_{p1} & \dots & a_{pk} \end{bmatrix}, \quad (2.1.4)$$

which is also called the factor loading matrix. Such a basis is not unique because the choice of Z_i , or the representation (2.1.3), is not unique. A basis is just meant to generate the entire space of factor variables which are linear combinations of some hypothetical variables introduced into the various

tests. From this point of view one basis is as good as any other. Of course, given one basis, orthogonal or oblique, the other can be derived by a linear transformation.

The choice of a suitable basis which is "psychologically meaningful" has largely rested with the psychologist, perhaps rightly so. But it is quite conceivable, once the psychological meaning is translated to mean some precisely stated restrictions on the basis, that its choice will turn out to be a problem of statistical estimation. In this sense the graphical methods of rotation of factor loadings advocated by Thurstone (17) and the quadrimax method of Neuhaus and Wrigley (13) are statistical methods of factor analysis, where the number of zero or small loadings is maximized. Such a restriction, or even the orthogonality of a basis, may not be the most helpful in leading to a suitable psychological interpretation or the discovery of "real entities." The choice of the restrictions to be imposed on the basis is perhaps a problem for psychological research. An objective method developed in this connection by Cattell (4, 5) seems to have interesting possibilities from a statistical point of view.

2.2. Principal Components

It was pointed out by Sir Cyril Burt that this method was originally put forward by Karl Pearson in 1901. But the statistical problems of estimation and testing connected with the principal components were first considered by Hotelling.

Hotelling (8) considers two types of problems:

First, without assuming a decomposition of the measurements as in (2.1.1), hypotheses are framed in terms of the latent roots of the correlation matrix with a view to studying the shape of the scatter of the standardized scores in the p -dimensional space or alternatively the relative importance of the different principal components in explaining the total variance. For instance, if some of the calculated roots are not significantly different then the components corresponding to them may be considered equally important.

Let us consider a specific hypothesis that the $(i+1)$ th to the $(i+r)$ th roots of the population correlation matrix (denoted by ρ) are equal. The value of i may be 0, 1, ... to $(p-r)$.

This hypothesis imposes a restriction on ρ , viz., that it admits the decomposition

$$\rho = \theta + \lambda I, \quad (2.2.1)$$

where θ is a matrix of rank $(p-r)$, λ is the common value of the roots from $(i+1)$ th to $(i+r)$ th, and I is the unit matrix.

Any such hypothesis or a similar one based on the dispersion matrix Σ instead of ρ can be tested by a likelihood-ratio criterion Λ , provided sample size is moderately large. Exactly how large the sample size should be is a

matter for further investigation. The statistic $(-2 \log_e \Lambda)$ is distributed in large samples as χ^2 with degrees of freedom equal to the number of restrictions on the free parameters imposed by the hypothesis.

The number of restrictions in a hypothesis of the form (2.2.1) is equal to the number of restrictions on the symmetric matrix θ with rank $(p - r)$ minus one for the unknown λ . Only $(p - r)$ rows and columns in θ are independent; the rest of the elements depend on them. Therefore, the number of restrictions on the elements of θ is

$$(p - r)(p - r + 1)/2,$$

and with one less we have

$$(p - r - 1)(p - r + 2)/2 \quad (2.2.2)$$

degrees of freedom for the χ^2 approximation.

If A is the estimated dispersion matrix from observations on n individuals, then the test criterion for the hypothesis (2.2.1) with Σ instead of ρ is

$$-(n - 1) \log \frac{|A|}{|\hat{\Sigma}|}, \quad (2.2.3)$$

where $|\hat{\Sigma}|$ is the estimated dispersion matrix under the conditions of the hypothesis. The latent roots of $\hat{\Sigma}$

$$\mu_1, \mu_2, \dots, \mu_i, \mu_{i+1}, \dots, \mu_{i+r}, \mu_{i+r+1}, \dots, \mu_p$$

are connected with the latent roots of A in the following way:

$$\lambda_j = \mu_j \quad (j = 1, \dots, i, i + r + 1, \dots, p),$$

$$\mu_{i+1} = \frac{\lambda_{i+1} + \dots + \lambda_{i+r}}{r}.$$

Since

$$|A| = \lambda_1 \dots \lambda_p, \quad |\hat{\Sigma}| = \mu_1 \dots \mu_p,$$

the ratio $|A|/|\hat{\Sigma}|$ is

$$\frac{\lambda_{i+1} \dots \lambda_{i+r}}{\left(\frac{\lambda_{i+1} + \dots + \lambda_{i+r}}{r} \right)^r}, \quad (2.2.4)$$

which is a suitable power of the ratio of the geometric to the arithmetic mean of the $(i + 1)$ th to $(i + r)$ th roots of A . From this point of view it would appear, by choosing $i = (p - r)$ in (2.2.4), that Bartlett's (1) test using the dispersion matrix instead of the correlations is valid for judging the significance of equality of the least r roots.

Unfortunately the test does not seem to reduce to the form (2.2.4) in terms of the roots of the observed correlation matrix R when the hypothesis is as stated in (2.2.1) in terms of the population correlation matrix ρ . The effect of standardizing the variables by the *sample standard deviations* is not properly allowed for by a criterion of the form (2.2.4). This is also partly revealed by Bartlett's own evaluation of the degrees of freedom by the expectation method in a simple case. They depend on the unknown correlations and reach the value (2.2.2) only in a limiting case, while for a genuine likelihood ratio this is not expected. The exact evaluation of the test criterion depends on complicated equations which require further investigation.

Secondly, Hotelling considers the problem of "testing the variances of components against the variance to be expected on account of the inaccuracy of the tests as revealed by their self-correlations or reliability coefficients." For this purpose a test score is thought of as made up of two parts, a true score with variance unity and a random error. Thus

$$x_i = X_i + \epsilon_i \quad (i = 1, \dots, p), \quad (2.2.5)$$

with the conditions $\text{cov}(\epsilon_i, \epsilon_j) = 0$, ($i \neq j$). The hypothesis stated above is interpreted to imply that the true scores X_i are linearly dependent, i.e., "the scatter diagram of the true scores will lie in a flat space of smaller dimensionality immersed in the p -dimensional space." If independent estimates of the variances of ϵ_i are available, either from an external source or by re-tests on individuals, there is no need to consider the true scores as random variables in order to test the above hypothesis. The general multivariate tests of dimensionality developed in more complicated situations are directly applicable for this problem. The non-stochastic model on the scores corrected for unreliabilities used in testing the second hypothesis provides a strong contrast to tests in factor analysis where, of necessity, all the variables (the common and specific factors) involved are considered to be stochastic, which makes the problem more complex.

3. *Special Characterizations of a Basis in Factor Analysis*

Using the vector notation

$$\underline{x} = (x_1, \dots, x_p), \quad \underline{z} = (z_1, \dots, z_p), \quad \underline{s} = (s_1, \dots, s_p).$$

The equation (2.1.1) can be written

$$\underline{x} = \underline{z} + \underline{s}. \quad (3.1)$$

The dispersion matrix of \underline{x} (using D for dispersion) is

$$D(\underline{x}) = D(\underline{z}) + D(\underline{s}),$$

or

$$\Sigma = \Gamma + \Delta, \quad (3.2)$$

where Σ , Γ and Δ are defined by equation (3.2). The covariance of \underline{z} and \underline{s} is zero because of conditions (2.1.2). The matrix Γ is positive semi-definite with rank $k < p$ and Δ is a positive-definite diagonal matrix. The equation (3.2) supplies the fundamental decomposition of the dispersion matrix Σ in terms of those of the hypothetical variables postulated by a factorial structure. If the rank of Γ is $k < p$, the space of common factors has a basis of k independent factors as shown in section 2.1. For a proper identification of the space and "an orderly selection of independent factors" there is a need to characterize a basis in a convenient way. A basis so characterized need not admit a psychological interpretation, for only mathematical and statistical convenience is being sought at this stage. A basis once obtained can always be transformed to meet other requirements. Two special characterizations are discussed here.

3.1 First Characterization

Let $\underline{l} = (l_1, \dots, l_p)$ be a vector of arbitrary coefficients giving rise to a new factor variable

$$\underline{lz}' = l_1 z_1 + \dots + l_p z_p.$$

The variation in the variable x_i explained by the factor variable \underline{lz}' , is

$$\frac{\text{cov}^2(x_i, \underline{lz}')}{V(\underline{lz}')} = \frac{(l_1 \gamma_{1i} + \dots + l_p \gamma_{pi})^2}{\underline{l} \Gamma \underline{l}'}, \quad (3.1.1)$$

assuming that $\underline{l} \Gamma \underline{l}'$, the variance of \underline{lz}' , is not zero. The total variation explained in all the variables is

$$\frac{\left\{ \sum_1^p (l_1 \gamma_{1i} + \dots + l_p \gamma_{pi})^2 \right\}}{\underline{l} \Gamma \underline{l}'} = \frac{\underline{l} \Gamma \underline{l}'}{\underline{l} \Gamma \underline{l}'}. \quad (3.1.2)$$

Let us choose \underline{l} such that (3.1.2) is a maximum. Differentiating with respect to the vector \underline{l} (see 14, p. 21), the equation leading to the optimum value λ of the ratio (3.1.2) and the vector $\underline{l} \Gamma$ is

$$\underline{l} \Gamma \Gamma - \lambda \underline{l} \Gamma = 0$$

or eliminating $\underline{l} \Gamma$

$$| \Gamma - \lambda I | = 0, \quad (3.1.3)$$

where I is the identity matrix. This shows that λ is the maximum latent root of Γ and $\underline{m} = \underline{l} \Gamma$ is the latent vector corresponding to it. Since the vector \underline{m} satisfies the equation

$$\underline{m} \Gamma = \lambda \underline{m},$$

and

$$\underline{m} = \underline{l} \Gamma,$$

so that

$$\underline{m}\Gamma = \lambda\underline{l}\Gamma, \quad (3.1.4)$$

the vector \underline{m} itself can be taken to be a solution of \underline{l} . We thus obtain the first factor variable as a linear combination of z_1, \dots, z_p . From the theory of canonical roots and vectors (14, p. 24), it would then follow that the second factor variable, which explains the highest proportion of the residual variation independently of the first, is the linear combination corresponding to the second canonical vector. There are as many linear combinations as there are non-zero roots λ , which is equal to the rank of the matrix Γ . The linear combinations of z_1, \dots, z_p supplied by the canonical vectors of zero roots of λ vanish identically, indicating the dependence of the factor variables associated with the measurements x_1, \dots, x_p .

The factor loading of the variable x_i on the first factor chosen above is the correlation between the two. The covariance is

$$\text{cov}(x_i, \underline{l}\underline{z}') = l_1\gamma_{1i} + \dots + l_p\gamma_{pi} = \lambda_1 l_i,$$

and if the variables x_i are initially chosen to have unit variance the correlation is

$$\frac{\lambda_1 l_i}{\sqrt{\underline{l}\Gamma\underline{l}'}} = \frac{\lambda_1 l_i}{\sqrt{\lambda_1 \underline{l}\underline{l}'}} = \frac{\sqrt{\lambda_1} l_i}{\sqrt{l_1^2 + \dots + l_p^2}}. \quad (3.1.5)$$

The factor loadings are then the elements of the first canonical vector suitably standardized. Similarly the factor loadings of any other factor are derived from the canonical vector defining the factor.

Even after exhausting all the independent factor variables, there still remains some variation left in \underline{x} to be explained by the specific factors unless the number of independent common factors is equal to p . In the problem originally considered by Hotelling, the successive components explaining variation in \underline{x} were not confined to the common factor portion \underline{z} but were also functions of the specific factors \underline{s} , which then are equivalent to linear functions of \underline{x} . Hotelling's principal components are, therefore, important in problems where the total variation of a measurement vector \underline{x} is sought to be accounted for, to the maximum amount possible, by a smaller number of linear functions of \underline{x} . The principal components of Hotelling are derived from the latent vectors of the matrix $\Sigma = \Gamma + \Delta$ instead of Γ alone as used above. It may be observed that when

$$\Delta = \delta^2 I,$$

i.e., when all the specific variables have the same variance δ^2 , a latent vector \underline{l} of $\Gamma + \Delta$ satisfying the equation

$$\underline{l}(\Gamma + \delta^2 I) = \mu \underline{l}$$

also satisfies the equation

$$\underline{l}\Gamma = (\mu - \delta^2)\underline{l} = \lambda\underline{l}$$

and is therefore a latent vector of Γ . The principal component analysis of Hotelling is thus a method of factor analysis with the factor loadings inflated keeping the same relative magnitudes, when all the specific variances are the same.

There is some arbitrariness in the above characterization of the basic set of factors because instead of maximizing the sum of the variations explained in x_1, \dots, x_p we could maximize a weighted sum and arrive at a different basis and consequently a different set of factor loadings. When the variables x_1, \dots, x_p are chosen to have unit variances the method adopted is equivalent to using reciprocals of total variances as weights.

The quantity δ_i^2 , the residual variance of x_i unexplained by the factor variables, satisfies the equation

$$\sigma_{ii} = \frac{\lambda_1 l_i^2}{\underline{l}\underline{l}'} + \frac{\lambda_2 m_i^2}{\underline{m}\underline{m}'} + \dots + \delta_i^2, \quad (3.1.6)$$

where $\underline{l}, \underline{m}, \dots$ are the latent vectors defining the factors $\underline{l}z', \underline{m}z', \dots$ and the subscript i relates to the i th element in the vectors.

The best formula for predicting a factor variable such as $\underline{l}z'$ from the observed measurements \underline{x} is obtained by the method of regression (16). If $\underline{k}\underline{x}'$ is the predicted value, then \underline{k} satisfies the equation

$$\underline{k}\Sigma = \underline{l}\Gamma, \quad \underline{k} = \underline{l}\Gamma\Sigma^{-1} = \lambda_1 \underline{l}\Sigma^{-1}, \quad (3.1.7)$$

and similarly for other factors. The characterization of the basis considered here together with methods of estimation is known as the principal factor analysis (PFA) (7).

3.2 Second Characterization

Instead of asking for a factor variable which explains as much of variation as possible of \underline{x} , we may pose the problem in a different way. What is that factor variable which is predictable from \underline{x} with the maximum possible precision? Or in other words, what is that factor variable which is maximally related to \underline{x} ? The solution to this problem depends on a canonical correlation analysis of the hypothetical factor variables \underline{z} with the measurable variables \underline{x} of which \underline{z} constitute a part.

If $\underline{l}z'$ and $\underline{q}\underline{x}'$ represent two linear combinations of factor variables and test scores, then according to the theory of canonical correlations (8) the correlation (or its square) between the two linear functions

$$\frac{(\underline{l}\Gamma\mathbf{q}')^2}{(\underline{l}\Gamma\underline{l}')(\mathbf{q}\Sigma\mathbf{q}')} \quad (3.2.1)$$

has to be maximized. Using the algebra developed in a similar genetic problem (14), the optimum value of the correlation ν is found to be a root of the equation

$$|\Gamma - \nu^2 \Sigma| = 0, \quad (3.2.2)$$

or

$$|\Sigma - \lambda \Delta| = 0, \quad \lambda = 1/(1 - \nu^2). \quad (3.2.3)$$

The vectors \underline{l} and \underline{q} are proportional and satisfy the same equation

$$\underline{l}(\Sigma - \lambda \Delta) = 0, \quad \underline{q}(\Sigma - \lambda \Delta) = 0. \quad (3.2.4)$$

That factor variable which is highly correlated with \underline{x} is $\underline{l}\underline{z}'$, where \underline{l} is the latent vector corresponding to the largest root of the determinantal equation (3.2.2). The second factor variable, uncorrelated with the first and possessing the highest correlation with \underline{x} is $\underline{m}\underline{z}'$, where \underline{m} is the latent vector corresponding to the second root of (3.2.2), and so on. We get as many factors as the number of non-zero values of ν^2 or values of λ greater than unity which is the same as the rank of Γ .

For any factor $\underline{l}\underline{z}'$ as determined above

$$\underline{l}\Gamma\underline{l}' = \nu^2 \underline{l}\Sigma\underline{l}' = \frac{\lambda - 1}{\lambda} \underline{l}\Sigma\underline{l}' = (\lambda - 1)\underline{l}\Delta\underline{l}'$$

$$\text{cov}(x_i, \underline{l}\underline{z}') = l_i \gamma_{1i} + \dots + l_p \gamma_{pi} = (\lambda - 1)l_i \delta_i^2.$$

The correlation between x_i and $\underline{l}\underline{z}'$,

$$\frac{(\lambda - 1)l_i \delta_i^2}{\sqrt{(\lambda - 1)\underline{l}\Delta\underline{l}'\sigma_{ii}}}, \quad (3.2.5)$$

is the factor loading of x_i on the factor $\underline{l}\underline{z}'$. This is again an element of \underline{l} multiplied by a constant. It can be shown that the same factor loadings are obtained if instead of (x_1, \dots, x_p) we consider $(c_1 x_1, \dots, c_p x_p)$ with the variables arbitrarily scaled. In the previous case it is necessary to reduce the variables (x_1, \dots, x_p) to unit standard deviation before proceeding to derive factors in order to achieve uniqueness of factor loadings.

To predict the factor measurements we use the regression equation as in (3.1.7). In this case it turns out that $\underline{l}\underline{z}'$, $\underline{m}\underline{z}'$, \dots , defined by the latent vectors of (3.2.3), can be best predicted by $\underline{l}\underline{x}'$, $\underline{m}\underline{x}'$, \dots , avoiding the complication of multiplication by Σ^{-1} necessary in the case of factors defined in the earlier characterization of the basic set (3.1.7).

The residual variance δ_i^2 in x_i , unexplained by the factor variables satisfies the equation

$$\sigma_{ii} = \frac{(\lambda_1 - 1)}{\underline{l}\Delta\underline{l}'} l_i^2 \delta_i^4 + \frac{(\lambda_2 - 1)}{\underline{m}\Delta\underline{m}'} m_i^2 \delta_i^4 + \dots + \delta_i^2, \quad (3.2.6)$$

similar to the formula (3.1.6) in the earlier case. This second characterization of a basis together with methods of estimation may be called canonical factor analysis (CFA) to bring out its connection with the theory of canonical correlations.

Factor analysis thus fits in a general theory of canonical correlations involving two sets of variables: one set being observable and the other set, observable as in multiple regression; dummy as in multiple discrimination; or hypothetical as in problems of genetic selection.

3.3 Which Is a Better Characterization?

This question is meaningless if we are dealing with the true values of the dispersion elements satisfying the conditions of a given rank of the factor-variable space because one can be transformed into the other, and in fact they may be replaced by any other basic set and they all serve the same purpose.

But this is no longer true when we have only *estimates* of the dispersion elements and factors are estimated by formally substituting for Σ the estimated quantities and choosing Δ to satisfy the equation (3.1.6) in the first case (PFA) and (3.2.6) in the second case (CFA). Which then is a better estimate of a basis?

From the point of view of statistical estimation, PFA gives a least-squares estimate (16, p. 119) and CFA, a maximum-likelihood estimate, when normality of the distribution of the observations is assumed. At present there is not much to choose between the two methods except for the following reasons. The maximum-likelihood estimation leads in general to better results when the distribution of the variables is specified. No suitable test based on the least-squares estimates is available while there exists an easily computable test criterion on the basis of maximum-likelihood estimates. No further computations are needed to obtain the factor measurements if the factors are estimated by CFA; in fact, in this method, factors are deduced from a description of their measurement.

There is another logical argument which may have to be borne in mind in deciding the issue. A rigorous hypothesis concerning the number of independent factor variables is perhaps never true, and a test of this null hypothesis can detect its falsehood only when there is a serious departure. If then by following a rule of behavior (as determined by a test criterion) we decide to extract a certain number of factors, any method of estimation may be looked upon as providing only a summary of all the factors in terms of a few dominant ones having a definite existence with magnitudes bigger than standard errors calculable from the observations. It is then of interest to examine whether one method of estimation leads to a better summary than the other and at the same time has low errors of estimation.

From this viewpoint, PFA may be thought of as providing the best k

(given number not necessarily exhaustive) factors explaining the maximum possible variance in the measurements while CFA, the best k factors which have in some sense highest possible correlations with the measurements. This may mean that while the first set attempts to explain as much as possible of the variations in the individual measurements, the latter set focuses on the correlations. Perhaps the psychological interest chiefly lies in the latter set, which offers a better explanation of the correlations between the measurements.

4. Estimation and Tests of Significance for Factors

4.1 Estimation of Factor Loadings

Let $A = (a_{ii})$ denote the observed dispersion matrix of the vector variable \underline{x} . This is sufficient for the estimation of Γ and Δ , the two components of the population dispersion matrix Σ . Following the equations (3.2.3, 3.2.4) of the second characterization, we have on substituting A for Σ

$$\begin{aligned} |A - \lambda\Delta| &= 0, \\ \underline{l}(A - \lambda\Delta) &= 0, \end{aligned} \quad (4.1.1)$$

where \underline{l} is a latent vector corresponding to the latent root λ . From the point of view of mechanical computations it is convenient to solve for

$$\underline{b} = \underline{l}\Delta^{1/2}, \quad \underline{b}\underline{b}' = 1, \quad (4.1.2)$$

in which case \underline{b} is the latent vector of

$$|\Delta^{-1/2}A\Delta^{-1/2} - \lambda I| = 0. \quad (4.1.3)$$

Let us suppose, for the sake of illustration, that we are extracting two factors. If $\underline{b} = (b_1, \dots, b_p)$ and $\underline{c} = (c_1, \dots, c_p)$ are the first two latent vectors of (4.1.3) corresponding to the roots λ_1 and λ_2 , then the equation (3.2.6) gives

$$a_{ii} = [(\lambda_1 - 1)b_i^2 + (\lambda_2 - 1)c_i^2 + 1]\delta_i^2, \quad (4.1.4)$$

or

$$\delta_i^2 = \frac{a_{ii}}{(\lambda_1 - 1)b_i^2 + (\lambda_2 - 1)c_i^2 + 1} = \frac{a_{ii}}{g_i^2}, \quad (4.1.5)$$

where g_i is defined by the last part of equation (4.1.5). The equation (4.1.3) can now be written in terms of the observed correlation matrix R instead of the dispersion matrix A

$$|GRG - \lambda I| = 0, \quad (4.1.6)$$

where the elements g_i of the diagonal matrix G satisfy

$$g_i = \sqrt{(\lambda_1 - 1)b_i^2 + (\lambda_2 - 1)c_i^2 + 1}. \quad (4.1.7)$$

The computational problem is then to solve for g_i 's satisfying the equations (4.1.6) and (4.1.7), where λ_1, λ_2 , are the latent roots of (4.1.6) and $\underline{b}, \underline{c}$, the latent vectors. A tentative method is to start with a trial matrix G and obtain successive approximations by solving (4.1.6) for λ_1, λ_2 , and $\underline{b}, \underline{c}$, and substituting in (4.1.7). The process is repeated until the g_i converge.

A better approximation to g_i is obtained by using the formula

$$g_i = \sqrt{\left(\frac{\lambda_1}{\lambda_e} - 1\right)b_i^2 + \left(\frac{\lambda_2}{\lambda_e} - 1\right)c_i^2} + 1, \quad (4.1.8)$$

where

$$\lambda_e = \frac{[\Sigma g_i^2]_0 - \lambda_1 - \lambda_2}{p - 2}, \quad (4.1.9)$$

the summation $[\Sigma g_i^2]_0$ refers to the g_i^2 at the previous stage used in equation (4.1.7) to obtain λ_1, λ_2 .

The two formulas (4.1.7) and (4.1.8) should agree towards the final stages when convergence is expected to be slow. But in the initial stages (4.1.8) may accelerate convergence.

The estimated factor loadings on the first and second factors at any stage of approximation are

$$\sqrt{\lambda_1 - 1} \underline{b}G^{-1}, \quad \sqrt{\lambda_2 - 1} \underline{c}G^{-1}.$$

The same method holds good for any number of factors. The estimates of factor loadings obtained from equations (4.1.6, 4.1.7) can be shown to satisfy the maximum-likelihood equations of Lawley (11, 12) and thus constitute one out of a number of possible solutions. The equations (4.1.6, 4.1.7) are in a proper shape to admit an iterative procedure for solution. The use of equation (4.1.7) seems to avoid a difficulty which may occur in the iterative procedures. The iterative method given by Lawley (16, p. 130) may suffer a breakdown on the initial iteration due to an improperly chosen trial set of factor loadings leading to imaginary values of the quantities commonly designated by " h_1, h_2, \dots ." This may also occur in PFA with the guessed communalities at the first stage.

4.2 Tests of Significance and Estimation of Number of Factors

It is also necessary to lay down some rules for determining the number of factors to be estimated. This is partially answered by any reasonable test for a specified number of factors. We determine that number of factors for which the chosen test does not show significance, while for any smaller number the hypothesis is contradicted. If the level of significance is based on the 5 per cent level, then this method leads us to a lower confidence limit to the number of factors. That is, we can assert, with a risk of only 5 per cent, that the number of factors is at least as large as that discovered by the above

procedure. This is no doubt an objective rule for determining the lower limit to the number of factors, but in practice it may be better to extract one or two more factors, depending on the magnitude of the residual roots. If one or two such roots are sufficiently bigger than unity (though not significantly so) it may be worth while to extract the factor corresponding to them also.

The hypothesis we propose to test is that the population dispersion matrix admits the decomposition

$$\Sigma = \Gamma + \Delta, \quad (4.2.1)$$

where Δ is a diagonal matrix with positive terms and Γ is a positive semi-definite matrix of rank $k < p$.

The test criterion we use is derived by the principle of likelihood ratio, assuming that the observations are normally distributed.

The exact distribution of the test criterion is not known but in large samples $(-2 \log)$ of the likelihood ratio is distributed as χ^2 with degrees of freedom equal to the number of independent restrictions on the elements of Σ imposed by the hypothesis (4.2.1). This hypothesis specifies the rank of the matrix $\Sigma - \Delta$ for suitably chosen Δ . If its rank is k , then by fixing the first k rows and columns the rest of the elements can be computed, which implies $(p - k)(p - k + 1)/2$ restrictions. Allowing for p unknown values in Δ , the number of restrictions is equal to

$$\frac{(p - k)(p - k + 1)}{2} - p = \frac{(p - k)^2 - p - k}{2}. \quad (4.2.2)$$

The test based on the likelihood-ratio criterion is

$$-(n - 1) \log \frac{|A|}{|\hat{\Sigma}|}, \quad (4.2.3)$$

where $\hat{\Sigma}$ is the estimated dispersion matrix using the maximum-likelihood equations of section 4.1. The multiplying coefficient $(n - 1)$, where n is the sample size, may be replaced by the more appropriate value for the χ^2 approximation to hold when n is not large,

$$\left(n - 1 - \frac{2p + 5}{6} - \frac{2k}{3} \right),$$

where p is the number of variables and k is the number of factors (1). Since the roots of the equation $|\Sigma - \lambda\Delta| = 0$ corresponding to k factors are estimated by $|A - \lambda\Delta| = 0$ or the equivalent forms (4.1.3), (4.1.6), it follows that the roots of $|\hat{\Sigma} - \lambda\Delta| = 0$ are

$$\lambda_1, \dots, \lambda_k, 1, 1, \dots, 1, \quad (4.2.4)$$

while the roots of $|A - \lambda\Delta|$ are, in descending order of magnitude,

$$\lambda_1, \dots, \lambda_k, \lambda_{k+1}, \dots, \lambda_p, \quad (4.2.5)$$

and, therefore,

$$\frac{|A|}{|\hat{\Sigma}|} = \lambda_{k+1} \cdots \lambda_p, \quad (4.2.6)$$

which is the product of the least $(p - k)$ roots, at the last stage of iteration of the equation $(4.1.6)$, $|GRG - \lambda I| = 0$.

The χ^2 test is

$$-(n - 1) \log (\lambda_{k+1} \cdots \lambda_p) \quad (4.2.7)$$

with $[(p - k)^2 - p - k]/2$ degrees of freedom apart from the slight refinement in the multiplying coefficient.

4.3 A Modified Criterion and Its Practical Use

It may be recalled that the likelihood-ratio criterion is the ratio of the maximum likelihood under the restrictions of the hypothesis (4.2.1) to that without any restrictions on Σ . It is of interest to examine how the ratio (or its logarithm) of the likelihoods is converging to the maximum value (or the negative of log ratio to its minimum value) during the iterative process. Fortunately this can be expressed in terms of the roots $\lambda_{k+1}, \dots, \lambda_p$ at any stage of the iterative process

$$-(n - 1)[\log (\lambda_{k+1} \cdots \lambda_p) - (p - k) \log \lambda_e], \quad (4.3.1)$$

where λ_e of (4.1.9) is the arithmetic mean of $\lambda_{k+1}, \dots, \lambda_p$. [Strangely the sequence (4.3.1) of statistics (likelihood ratios), which converge ultimately to the test criterion (maximum-likelihood ratio), resembles Bartlett's (1) ratio test but, of course, the roots λ_i are obtained differently and the ratios are used with different degrees of freedom. From this analysis it would appear that Bartlett's ratio is an initial approximation to the actual test criterion.] (4.3.1) converges to

$$-(n - 1) \log (\lambda_{k+1} \cdots \lambda_p)$$

at the final stage when

$$(p - k) = \lambda_{k+1} + \cdots + \lambda_p. \quad (4.3.2)$$

Suppose that (4.3.1) is not significant as χ^2 with $[(p - k)^2 - p - k]/2$ degrees of freedom, at any stage, then the same conclusion is reached even after completing the iterative process. If a test of significance is the only aim of analysis, then, sometimes, iteration can be stopped at some stage. Even if the result is significant, it is possible to terminate the computations provided the change in (4.3.1) at that stage is small from one cycle of operations to another.

The modified criterion is extremely useful in practice when the object

of the analysis is to estimate the number of factors (lower confidence value) as well as the factor loadings.

Before proceeding with the cycle of operations for estimation, let us fix some high value of k as the number of factors and calculate the roots after one or two iterations. At this stage, find that value of r for which Λ_r (with d.f. $[(p - r)^2 - p - r]/2$) is not significant, but Λ_{r-1} is. This shows that the number of factors is not greater than r . We may set the number of factors provisionally at r and continue the process of estimation. Each time we may calculate Λ_{r-1} and Λ_r to see whether Λ_{r-1} becomes not significant at any stage. If it is not significant, there is a case for switching over to $(r - 1)$ factors instead of r .

5. Summary

The experimental situation and the nature of the data on which the technique of factor analysis can be successfully employed may be stated as follows. Each of the p measurements on an individual has a linear regression on a common set of a few hypothetical variables or factors: The deviations from regression for any two measurements are uncorrelated. The factor analysis seeks the smallest number of independent hypothetical variables necessary to explain the intercorrelations between the measurements.

If R is the observed correlation matrix, the computational problem of factor analysis depends on the solution of the diagonal matrix G satisfying the equations

$$|GRG - \lambda I| = 0, \quad (5.1)$$

$$g_i = [(\lambda_1 - 1)a_{1i}^2 + \dots + (\lambda_k - 1)a_{ki}^2 + 1]^{1/2}, \quad (5.2)$$

where k is the number of factors assumed, $\lambda_1, \dots, \lambda_k$, are the first k largest roots of (5.1) and $a_i = (a_{i1}, \dots, a_{ip})$ is the latent vector corresponding to the root λ_i . Once G is found to satisfy the equations (5.1, 5.2), then the factor loadings are given by

$$(\lambda_j - 1)a_j G^{-1} \quad (j = 1, \dots, k),$$

and the test of the hypothesis that k factors are adequate to explain the intercorrelations is

$$\chi^2 = -(n - 1) \log_e (\lambda_{k+1} \dots \lambda_p)$$

with $[(p - k)^2 - p - k]/2$ degrees of freedom. The lower confidence limit to the number of factors is the smallest value of k for which χ^2 is not significant.

Some research remains to be done to find an elegant computational technique for solving the equations (5.1, 5.2). The method available at present is to guess suitable values of g_i , substitute in (5.1) and obtain better

approximations to g_i by using (5.2). This process is continued until convergence is secured. Unfortunately this appears to be a slow process unless the initial values of g_i are very near the true values. Even with a good set of trial values the problem can be best tackled only on an electronic computer when large numbers of variables are involved. A suitable program for Illiac is being written by Mr. Golub of the Digital Computer Laboratory at the University of Illinois. A numerical example solved on a tentative program is reported below. Full details will be presented soon.

First it may be noted that the relation between g_i and the communality h_i^2 for the i th variate is

$$g_i = 1/\sqrt{1 - h_i^2},$$

so that good trial values of g_i are available once the communalities are approximately determined by an initial factorization of the correlation matrix by a simpler method, such as the centroid. Another method suggested in the literature is to choose the squared multiple correlation as an estimate of the communality. In many cases it is sufficient to start with the initial approximation $g_i = 1/2$.

Second, although the test involves the product of the roots at the final stages of convergence, it is useful to compute at intermediate stages the statistic

$$\chi^2 = -(n-1)[\log_e(\lambda_{k+1} \cdots \lambda_p) - (p-k) \log_e(\lambda_{k+1} + \cdots + \lambda_p)],$$

which, when not significant, implies the nonsignificance of the ultimate χ^2 . We could stop at any stage after this, provided further iterations do not considerably alter the factor loadings.

The following correlation matrix was presented by Davis (6) in an attempt to study factors of comprehension in reading.

1.00								
.72	1.00							
.41	.34	1.00						
.28	.36	.16	1.00					
.52	.53	.34	.30	1.00				
.71	.71	.43	.36	.64	1.00			
.68	.68	.42	.35	.55	.76	1.00		
.51	.52	.28	.29	.45	.57	.59	1.00	
.68	.68	.41	.36	.55	.76	.68	.58	1.00

Assuming a single factor, the χ^2 was calculated and found to be significant. This indicated more than one factor. Under the hypothesis of two factors the value of χ^2

$$-(n-1)[\log(\lambda_3 \cdots \lambda_9) - (9-2) \log(\lambda_3 + \cdots + \lambda_9)]$$

came down to 29.73 at an early stage of iteration. This being less than 30.1, the 5 per cent significance value of χ^2 with 19 degrees of freedom, the hypothesis of two factors stands unrejected. So the data admit an interpretation in terms of two significant factors only. A fairly stabilized set of factor loadings are

Factor 1	.845	.817	.477	.401	.669	.891	.834	.651	.833
Factor 2	-.309	-.084	.012	.153	.161	.145	.081	.122	.080

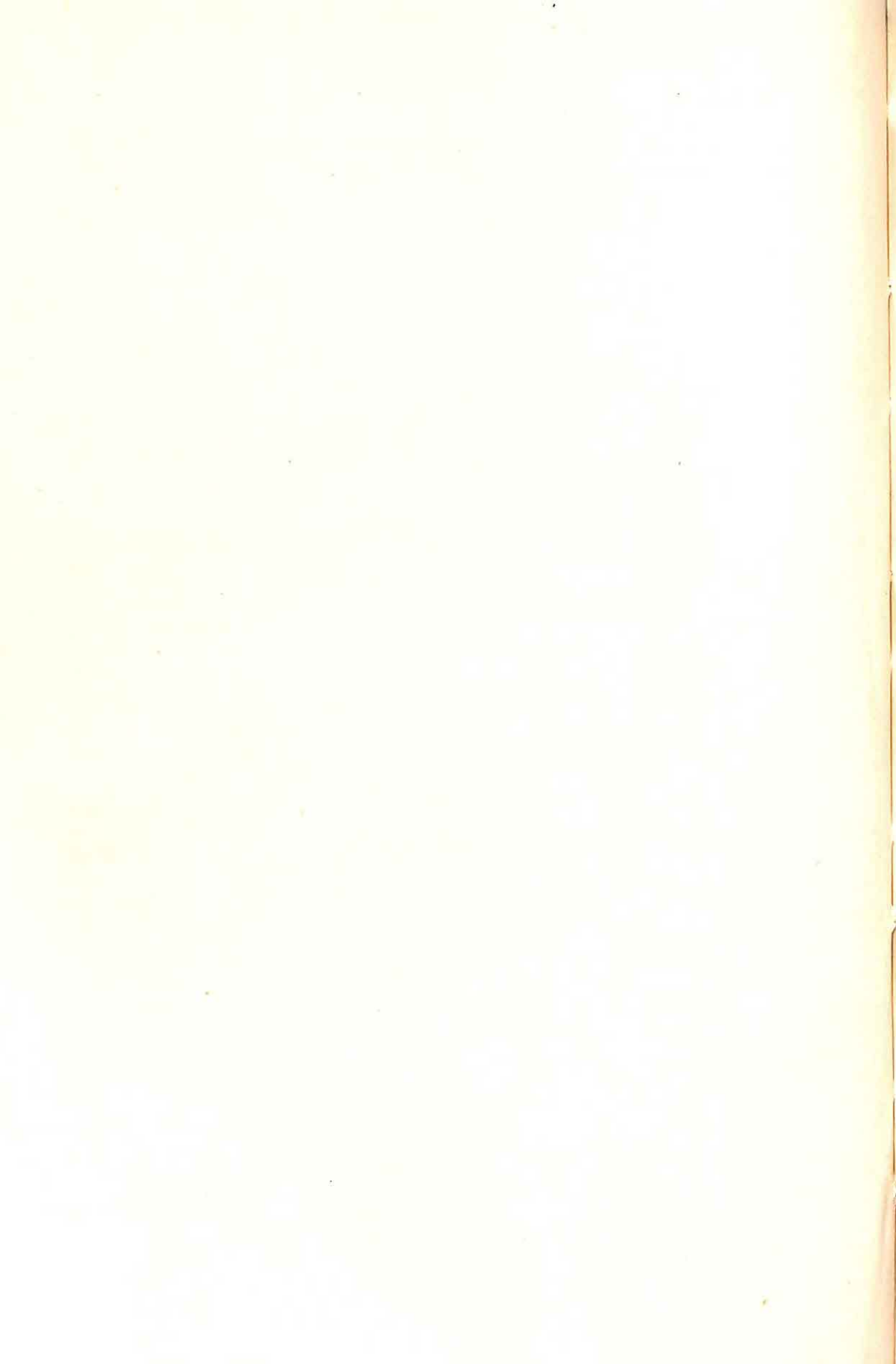
I wish to thank Dr. C. F. Wrigley, who read the manuscript and offered some helpful comments.

REFERENCES

1. Bartlett, M. S. Tests of significance in factor analysis. *Brit. J. Psychol., Statist. Sect.*, 1950, 3, 77-85.
2. Bartlett, M. S. A further note on tests of significance in factor analysis. *Brit. J. Psychol., Statist. Sect.*, 1951, 4, 1-2.
3. Burt, C. Tests of significance in factor analysis. *Brit. J. Psychol., Statist. Sect.*, 1952, 5, 109-133.
4. Cattell, R. B. Parallel proportional profiles. *Psychometrika*, 1944, 9, 267-283.
5. Cattell, R. B. The description and measurement of personality. Yonkers, New York: World Book Co., 1946.
6. Davis, F. B. Fundamental factors of comprehension in reading. *Psychometrika*, 1944, 9, 185.
7. Holzinger, K. J., and Harman, H. H. Factor analysis. Chicago: Univ. Chicago Press, 1941.
8. Hotelling, H. Analysis of a complex of variables into principal components. *J. educ. Psychol.*, 1933, 24, 417-441, 498-520.
9. Hotelling, H. Relations between two sets of variates. *Biometrika*, 1936, 28, 321-377.
10. Kendall, M. G. Factor analysis. *J. roy. stat. Soc., Series B*, 1950, 12, 60.
11. Lawley, D. N. The estimation of factor loadings by the method of maximum likelihood. *Proc. roy. Soc. Edin.*, 1940, 60, 64-82.
12. Lawley, D. N. Further investigations in factor estimation. *Proc. roy. Soc. Edin.*, 1941, 61, 176-185.
13. Neuhaus, J. O., and Wrigley, C. F. The quadrimax method: an analytic approach to orthogonal simple structure. Manuscript on file in the Univ. Illinois Library, 1953.
14. Rao, C. R. Advanced statistical methods in biometric research. New York: Wiley, 1952.
15. Rao, C. R. Discriminant functions for genetic differentiation and selection. *Sankhyā*, 1953, 12, 229.
16. Thomson, G. H. The factorial analysis of human ability. London: Univ. London Press 5th ed., 1951.
17. Thurstone, L. L. A new rotational method in factor analysis. *Psychometrika*, 1938, 3, 199-218.

Manuscript received 2/19/54

Revised manuscript received 5/11/54



RELIABILITY FORMULAS FOR NONCOMPLETED OR SPEEDED TESTS*

LOUIS GUTTMAN

THE ISRAEL INSTITUTE OF APPLIED SOCIAL RESEARCH

New formulas are developed to give lower bounds to the reliability of a test, whether or not all respondents attempt all items. The formulas apply in particular, then, to completed tests, pure speed tests, pure power tests, and any mixture of speed and power. For the case of completed tests, the formulas give the same answer as certain standard ones; for noncompleted tests the formulas give a correct answer where previous standard formulas are inappropriate. The formulas hold both in the sense of retest reliability and of parallel tests.

I. Introduction

Recently, there has been increasing awareness of an important inadequacy of all standard formulas that are in current use for studying reliability of tests. These formulas are not appropriate for tests in which all items are not attempted by everybody. In particular, they do not hold for speeded tests (cf. 1).

The present paper proposes a new analysis of the problem, and provides some practical formulas that hold whether or not the tests are completed. The case of completed tests emerges as a specialization of the present analysis. Thus, formulas developed here hold for pure speed tests, pure power tests, and for tests which are partly speed and partly power.

An important example of one of the practical formulas developed here is as follows. Consider a test composed of m dichotomous items. Each item is scored *unity* if answered correctly, *zero* if answered incorrectly or not attempted. Each person's total score is the sum of his scores on the m items. Suppose the test is administered once to a large population of individuals and that there are $m - n$ items each of which has *zero variance* in its scores; that is, on each of these $m - n$ items either all people scored 0 or all scored 1. In particular, all items not attempted by anybody are in this subset of $m - n$. The n items with *positive variance* will have their statistics on the single trial denoted as follows:

x_j = proportion of the population that answered the j th item correctly
($j = 1, 2, \dots, n$);
 p_j = proportion of the population that attempted the j th item
($j = 1, 2, \dots, n$), regardless of whether the answer was correct or not.

*This research was facilitated by an uncommitted grant-in-aid to the writer from the Behavioral Sciences Division of the Ford Foundation.

Let s_i^2 denote the variance of total scores on the trial. It is immaterial whether s_i^2 is computed from all m items or only from the n of positive variance, since adding items with zero variance will not change s_i^2 . Let D' be defined by

$$D' = \sum_{j=1}^n (n-j) \sqrt{x_j(1-p_j)}, \quad (1)$$

and let L'_3 be defined as

$$L'_3 = \frac{n}{n-1} \left(1 - \frac{D' + \sum_{j=1}^n x_j(1-x_j)}{s_i^2} \right). \quad (2)$$

Then, if ρ_i^2 denotes the reliability coefficient for the total test scores, we prove below that L'_3 is a lower bound to ρ_i^2 , i.e.,

$$L'_3 \leq \rho_i^2 \leq 1. \quad (3)$$

Another lower bound to ρ_i^2 derived below can be designated by L''_3 . To compute it, first compute D'' by

$$D'' = 2 \sum_{j=1}^n \left(\sqrt{1-p_j} \sum_{g=j+1}^n \sqrt{x_g} \right), \quad (4)$$

and then use D'' in place of D' in (2).

Sometimes L''_3 may be better than L'_3 , depending on whether $D'' < D'$ or not. A third lower bound, and one that is always better than either L'_3 or L''_3 , is given by inequality (52) below. Directions which can lead to even better lower bounds are indicated at the end of this paper. Since D' and D'' can be relatively substantial numbers compared to s_i^2 in noncompleted—especially speeded—tests, they can yield very low bounds L'_3 and L''_3 —even negative (or useless) ones. In part, this can be due to the greater room for unreliability for noncompleted tests as compared with completed tests, and in part to a certain loss of information that occurs in the derivation of our present formulas. This loss can sometimes be made up by devising more specialized formulas for special cases. Our present formulas are for a very wide class of cases, and hence cannot be most efficient for every subclass separately.

In (3), ρ_i^2 can be interpreted from the point of view either of retest reliability or of parallel tests; the same lower bound (2) ensues in each case (cf. 3). The same freedom of interpretation holds for all the formulas in the present paper, since we restrict ourselves to but a single trial for the actual numerical computations.

To establish that L'_3 is a lower bound to ρ_i^2 , we begin by using the results of a previous paper (3), wherein some important fundamental tautologies and inequalities for ρ_i^2 are developed that make no assumptions whatsoever.

For practical use, a certain quantity denoted there by δ must be observable, or at least bounded from above. The contribution of the present paper is essentially to establish upper bounds to δ that are observable from a single trial, given a certain assumption discussed below. The quantities D' and D'' defined in (1) and (4) are such bounds to δ for the type of test described above, and L'_3 and L''_3 are the resulting modifications (as computed from a single trial) of the lower bound to ρ_i^2 denoted by λ_3^* in (3).

Notice that if everybody attempts all items in the test just discussed, so that $p_i = 1$ ($j = 1, 2, \dots, n$), then the radicals in the right of (1) and (4) vanish for all j , making $D' = D'' = 0$. In such a case, according to (2), both L'_3 and L''_3 become the same as the lower bound L_3 discussed in (2), or the usual lower bound for the case of completed tests wherein all items are experimentally independent.

$$L'_3 = L''_3 = L_3 = \frac{n}{n-1} \left(1 - \frac{\sum_{i=1}^n x_i(1-x_i)}{s_i^2} \right) \quad (D' = D'' = 0). \quad (5)$$

[As has been pointed out in (2), L_3 is algebraically the same as formulas deduced in other contexts by Kuder and Richardson and by Hoyt, but its derivation, interpretation, and use in (2) are quite different from those of the others by virtue of the differing contexts. The context in which L_3 was originally derived is a special case of the context of L'_3 and L''_3 or of the present paper. It is not clear at present whether the Kuder-Richardson or the Hoyt formulations can be readily extended to the problem of noncompleted or speeded tests.] L'_3 and L''_3 are more general than L_3 in that they allow for possible experimental dependence among the items due to noncompletion of the test.

Other lower bounds to ρ_i^2 which allow for possible experimental dependence among items are also developed here, as well as bounds for tests in which the items are not dichotomies, or are not scored with 0-1 weights.

As one of the referees of this paper has pointed out, it may be desirable also to estimate the total error variance itself and not just ρ_i^2 . The variance in question is that denoted by ϵ^2 in 3, p. 229. Error variances will in general vary from individual to individual, especially in speeded tests, and ϵ^2 is their mean over all individuals. The lower bounds to ρ_i^2 are easily converted into upper bounds for ϵ^2 by virtue of the relationship: $\epsilon^2 = \sigma_i^2(1 - \rho_i^2)$, where σ_i^2 is estimated by s_i^2 . Thus $\epsilon^2 \leq s_i^2(1 - L)$, where L is any lower bound to ρ_i^2 .

II. Notation

For the proofs, the notation of the previous paper (3) will be followed. A slight modification here is that m denotes the number of items, or part-scores, in the test, while n denotes the number of part-scores with *observed* variance greater than zero. Only these n actually variable part-scores affect

the reliability of the total scores, and we get more efficient lower bounds by using n in place of the larger number m .

We consider here only the case where $n \geq 2$, since we wish largely to restrict ourselves to *information about reliability that is obtainable from an internal analysis of but one trial* of the test. This requires that the test have at least two subscores ($m \geq 2$), and in particular that at least two subscores each have a variance greater than zero ($n \geq 2$). In what follows, we assume that items of the test with positive observed variance are the only ones being considered.

Let x_{ijk} be the score of person i on the j th item of the test (with positive variance) on trial k ($j = 1, 2, \dots, n$), and let t_{ik} be the sum of the n part-scores

$$t_{ik} = \sum_{j=1}^n x_{ijk} . \quad (6)$$

The scoring scheme for any item can be arbitrary, except for the scores to be given to nonattempted items. We shall assume that a nonattempted item is given a score no higher than the lowest possible score for that item when attempted. More specifically, we shall assume that no negative scores are given to any item, and that nonattempted items are all scored zero. Should a scoring scheme originally allow for negative scores, it can easily be converted into the non-negative form we require by addition of a suitable constant to each part. This will not change the reliability coefficient ρ_i^2 in any way, nor any of the variances required in our formulas. A non-negative scoring scheme will yield total scores that correlate perfectly with those from the original scheme from which it is derived by this adding of constants.

It should be clear that we are excluding from our present analysis the case where an incorrect answer to an item is scored lower than omitting that item.

Let m_j be the maximum score obtainable on the j th item. Our assumption is then that the scoring scheme is in such a form that for all i, j , and k

$$0 \leq x_{ijk} \leq m_j , \quad (7)$$

and in particular that $x_{ijk} = 0$ if person i omits item j on trial k .

The population of persons and the universe of trials will both be assumed to be indefinitely large in order to avoid discussion here of sampling problems. Ultimately, only one trial from the universe need be made in practice to provide empirical data for use in our lower bounds.

The expected values over the trials of the x_{ijk} are denoted by X_{ij} ,

$$X_{ij} = E_k x_{ijk} . \quad (8)$$

The error of unreliability—or the experimental error—for person i on item j in trial k is $x_{ijk} - X_{ij}$. The covariance over trials between two items j and g is defined separately for each individual i and is denoted by $\gamma_{x_{ij}x_{ig}}$,

$$\gamma_{x_{ij}x_{ig}} = E_{\substack{k \\ g \neq j}} x_{ijk}x_{igk} - X_{ij}X_{ig} . \quad (9)$$

It has been shown how the reliability coefficient ρ_i^2 of the total test score, as well as lower bounds to ρ_i^2 , depend directly on the quantity δ defined from the covariances in (9) by

$$\delta = \sum_{\substack{g \neq j \\ i}} E \gamma_{x_{ij}x_{ig}} . \quad (10)$$

Another way of writing the right member of (10), which is more convenient for our purposes, is

$$\delta = 2 \sum_{j=1}^n \sum_{g=j+1}^n E \gamma_{x_{ij}x_{ig}} . \quad (11)$$

The right member of (11) equals the right member of (10) by virtue of the fact that, from (9), $\gamma_{x_{ig}x_{ij}} = \gamma_{x_{ij}x_{ig}}$, or these covariances are symmetric in g and j .

If part scores g and j are experimentally independent (that is, statistically independent over trials) for the i th person, then $\gamma_{x_{ij}x_{ig}} = 0$. If these two part scores are independent for all persons in the population, then $E \gamma_{x_{ij}x_{ig}} = 0$.

The converse need not hold, of course. Since nonzero covariances can be either positive or negative, we can have $E \gamma_{x_{ig}x_{ij}} = 0$ by having positive covariances for some people and negative ones for others.

Similarly, if *all items* are statistically independent for *all people*, then δ must vanish. However, we can have $\delta = 0$ even though not all items are statistically independent for all people, for again positive and negative covariances within and/or between pairs of items can cancel each other.

The role δ plays in lower bounds to ρ_i^2 is illustrated by the third universal lower bound, λ_3^* , developed in (3),

$$\lambda_3^* = \frac{n}{n-1} \left(1 - \frac{\delta + \sum_{j=1}^n \sigma_{x_j}^2}{\sigma_t^2} \right) . \quad (12)$$

The variances in the right member of (12) are defined as follows. The notation for the expected ("true") individual total scores on the test is

$$T_i = E_k t_{ik} . \quad (13)$$

The respective over-all means for part and total scores are

$$\xi_j = E_i X_{ij} , \quad \tau = E_i T_i . \quad (14)$$

Then,

$$\sigma_{x_i}^2 = E_i E_k (x_{ijk} - \xi_j)^2 , \quad \sigma_t^2 = E_i E_k (t_{ik} - \tau)^2 . \quad (15)$$

Each of the four types of parameters defined in (14) and (15) is observable on a single trial. For any fixed value of k , consider only expectations over i :

$$E_i x_{ijk}, \quad E_i t_{ik}, \quad E_i (x_{ijk} - E_i x_{ijk})^2, \quad E_i (t_{ik} - E_i t_{ik})^2. \quad (16)$$

It has been shown (2) how the variance of each of these quantities is zero over trials, the basic assumption being that each respondent answers independently of the other respondents (that there is no cribbing one from the other, etc.). Thus, the probability is unity that in any given trial k , the four quantities in (16) are respectively equal to

$$\xi_i, \quad \tau, \quad \sigma_{x_i}^2, \quad \sigma_i^2. \quad (17)$$

If the population of respondents is not infinite, or if only a finite sample is drawn from an infinite population, then the variances of the quantities in (16) will be positive, and the quantities will be only estimates of the respective quantities in (17). As in the previous papers, we shall not be concerned here with sampling error and assume for convenience that the operators E and E_k are always over an infinite population or universe.

III. Some Basic Identities

We need some further notation to study what happens when items are not attempted. Let

$$p_{ijk} = \begin{cases} 1 & \text{if person } i \text{ attempts item } j \text{ on trial } k \\ 0 & \text{otherwise.} \end{cases} \quad (18)$$

Furthermore, let

$$P_{ij} = E_k p_{ijk}, \quad \pi_i = E_i P_{ij}. \quad (19)$$

Thus, P_{ij} is the *proportion of the trials* in which person i attempts the item j and π_i is the mean of such proportions over all individuals. Just like the quantities in (17), π_i is observable from a single trial, namely by computing $E_i p_{ijk}$, or the proportion attempting the item in the given trial. The proof of this observability is of exactly the same nature for the parameters in (17) (cf. 2).

The universe of trials is thus divided into two sub-universes for each person and for each item: those trials in which he attempts the item (so that $p_{ijk} = 1$) and those in which he does not attempt the item (so that $p_{ijk} = x_{ijk} = 0$).

Given notation (18), the following simple and important identity holds:

$$p_{ijk} x_{ijk} = x_{ijk}. \quad (20)$$

For if $p_{ijk} = 0$, the x_{ijk} is zero by our convention that unattempted items are scored zero. And if $p_{ijk} = 1$, then (20) holds by direct multiplication.

Further notation needed refers to certain expected values of the x_{ijk} or x_{igk} . Let $X_{ig}^{(i)}$ be the expected value of x_{igk} over that subset of trials in which person i attempts item j (or for which $p_{ijk} = 1$),

$$X_{ig}^{(i)} = E_k p_{ijk} x_{igk} / P_{ij} . \quad (21)$$

Stating (21) for the case where $g = j$, using (20), and remembering (8) yield

$$X_{ij}^{(i)} = X_{ij} / P_{ij} . \quad (22)$$

Finally, we need a basic identity relating two covariances. Let $\gamma_{x_{ij}x_{ig}}^{(i)}$ denote the covariance between errors of unreliability, analogous to (9), for person i on items j and g , but only over the sub-universe of trials in which person i attempts item j (or where $p_{ijk} = 1$),

$$\gamma_{x_{ij}x_{ig}}^{(i)} = \frac{E_k p_{ijk} x_{ijk} x_{igk}}{P_{ij}} - X_{ij}^{(i)} X_{ig}^{(i)} . \quad (23)$$

Using (20) and (22) in (23) and then multiplying through by P_{ij} show that

$$P_{ij} \gamma_{x_{ij}x_{ig}}^{(i)} = E_k x_{ijk} x_{igk} - X_{ij} X_{ig}^{(i)} . \quad (24)$$

Then, from (9) and (24),

$$\gamma_{x_{ij}x_{ig}} = P_{ij} \gamma_{x_{ij}x_{ig}}^{(i)} + X_{ij} (X_{ig}^{(i)} - X_{ig}) . \quad (25)$$

Identity (25) is our basic tool for examining the dependence among experimental errors due to noncompletion of tests. It breaks the over-all covariance between errors, $\gamma_{x_{ij}x_{ig}}$, into two component parts, as expressed by the right member.

IV. A Basic Assumption and Its Consequences

Up until now, we have derived only identities or tautologies which are universally true, given a non-negative scoring scheme. The basic assumption from now on is that, if person i attempts item j , then his score on any later item g ($g > j$) will be experimentally independent of his score on this attempted item j . That is, we are considering here the case where dependence is due solely to *omissions*, so that if a part is attempted, no further experimental dependence holds. This may be true, for example, of pure speed tests as well as many other tests, including some pure power tests where omissions may be scattered and not consecutive. The experimental dependence between items discussed in (1) is in particular true of pure speed tests: if a person does not reach a certain item, he certainly will not reach later items on the same trial; thus the dependence is due to *nonattempts* or omissions. We

shall assume that *attempted* items do *not* lead to experimental dependence but only to dependence among true or expected scores. In particular, we assume *error* covariances of the following type to vanish:

$$\gamma_{x_{ij}x_{ig}}^{(i)} = 0 \quad (g > j). \quad (26)$$

If hypothesis (26) is true, then (25) reduces to

$$\gamma_{x_{ij}x_{ig}} = X_{ij}(X_{ig}^{(i)} - X_{iv}) \quad (g > j). \quad (27)$$

According to (27), the total dependence over trials between item scores is a function of the three expected values in the right member. Should $X_{ig}^{(i)}$ equal X_{iv} , or the fact that the item j is attempted does not change the expected value on the (later) item g , then we would have $\gamma_{x_{ij}x_{ig}} = 0$ according to (27), or we would be back to what is assumed (implicitly or explicitly) in all previous standard reliability formulas. But if $X_{ig}^{(i)} \neq X_{iv}$, then experimental dependence must hold between items j and g for person i .

We now wish to establish a useful upper bound to $E_i \gamma_{x_{ij}x_{ig}}$. From (21), since $p_{ijk} \leq 1$ and all quantities involved are non-negative, we see that

$$X_{ig}^{(i)} \leq X_{ig}/P_{ii}. \quad (28)$$

Using (28) in (27), and remembering (22),

$$\gamma_{x_{ij}x_{ig}} \leq X_{ij}^{(i)} X_{ig}(1 - P_{ii}) \quad (g > j). \quad (29)$$

From (7) and (21),

$$X_{ig}^{(i)} \leq m_g. \quad (30)$$

Setting $g = j$ in (30) and using the result in (29) yield

$$\gamma_{x_{ij}x_{ig}} \leq m_j X_{ig}(1 - P_{ii}) \quad (g > j). \quad (31)$$

Taking expected values of both members of (31) over i yields

$$E_i \gamma_{x_{ij}x_{ig}} \leq m_j E_i X_{ig}(1 - P_{ii}) \quad (g > j). \quad (32)$$

Now, from Schwarz' inequality,

$$E_i X_{ig}(1 - P_{ii}) \leq \sqrt{E_i X_{ig}^2 E_i (1 - P_{ii})^2}. \quad (33)$$

Also, from (7) (written with g in place of j),

$$X_{ig}^2 \leq m_g X_{ig}, \quad (34)$$

and from the fact that $0 \leq P_{ii} \leq 1$,

$$(1 - P_{ii})^2 \leq 1 - P_{ii}. \quad (35)$$

Using (35) and (34) in (33), and then the result in (32) [remembering (14)] produces the desired inequality

$$E \gamma_{x_i x_{ig}} \leq m_i \sqrt{m_g \xi_g (1 - \pi_i)} \quad (g > j). \quad (36)$$

In (36), both π_i and ξ_g , are observable from a single trial.

The upper bound to δ that we are seeking is obtained from (36) by summing both members over g and j (except for $g = j$) according to (11),

$$\delta \leq 2 \sum_{j=1}^n \left(m_j \sqrt{1 - \pi_j} \sum_{g=j+1}^n \sqrt{m_g \xi_g} \right). \quad (37)$$

In a test composed only of dichotomous items, each of which is scored zero or unity, we have $m_i \equiv 1$. For such a test, (37) reduces to

$$\delta \leq 2 \sum_{j=1}^n \left(\sqrt{1 - \pi_j} \sum_{g=j+1}^n \sqrt{\xi_g} \right) \quad (m_i \leq 1). \quad (38)$$

Inequality (38) holds for $m_i < 1$ as well as for $m_i = 1$, and we have so stated it; this is the formula given in 3, p. 64. What we have defined above as D'' in (4) is the right member of (38) as computed from a single trial.

Reviewing the proof shows that (37) holds for a much less restrictive hypothesis than (26). We can assume the inequality

$$\gamma_{x_i x_{ig}}^{(i)} \leq 0 \quad (g > j) \quad (39)$$

in place of only the equality of (26), and again arrive at (37). Under what circumstances an actually negative error covariance can be justifiably hypothesized remains a problem to be explored.

V. Another Upper Bound for δ

Using assumption (26), or (39), it is possible to arrive at other useful inequalities for $E \gamma_{x_i x_{ig}}$ and for δ in place of (36) and (37). For example, the following inequality will be established:

$$E \gamma_{x_i x_{ig}} \leq \frac{1}{2} m_g \sqrt{m_i \xi_i (1 - \pi_i)} \quad (g > j). \quad (40)$$

For the proof of (40), use (30) in (27) to obtain

$$\gamma_{x_i x_{ig}} \leq X_{ii} (m_g - X_{ig}) \quad (g > j). \quad (41)$$

Since the right members of (41) and (31) are always non-negative, their geometric mean is never smaller than the smaller of the two, so we can write from (41) and (31) that

$$\gamma_{x_i x_{ig}} \leq \sqrt{m_i X_{ii} (1 - P_{ii}) X_{ig} (m_g - X_{ig})} \quad (g > j). \quad (42)$$

Notice that the left member of (42) may be negative, or that (42) does not refer to the absolute value of $\gamma_{x_i x_{ig}}$. Now, the quantity $X_{ig} (m_g - X_{ig})$,

when regarded as a function of X_{ig} , reaches a maximum when $X_{ig} = m_g/2$, so we always have

$$X_{ig}(m_g - X_{ig}) \leq m_g^2/4. \quad (43)$$

Using (43) in (42) yields

$$\gamma_{x_{ij}x_{ig}} \leq \frac{1}{2}m_g \sqrt{m_i X_{ij}(1 - P_{ij})} \quad (g > j). \quad (44)$$

From Schwarz' inequality, and then notation (14) and (19),

$$E_i \sqrt{X_{ij}(1 - P_{ij})} \leq \sqrt{\xi_i(1 - \pi_i)}. \quad (45)$$

Taking expectations over i of both members of (44) and then using (45) yield the desired inequality (40).

Summing both members of (40) over g and j , remembering (11), yields another upper bound to δ ,

$$\delta \leq \sum_{j=1}^n \left(\sqrt{m_i \xi_i(1 - \pi_i)} \sum_{g=j+1}^n m_g \right). \quad (46)$$

For the special case of a test composed of dichotomous items scored 0 or 1, or where $m_i \equiv 1$, we have

$$\sum_{g=j+1}^n m_g = n - j \quad (m_g \equiv 1), \quad (47)$$

so that for this special case (46) reduces to

$$\delta \leq \sum_{j=1}^n (n - j) \sqrt{\xi_i(1 - \pi_i)} \quad (m_i \equiv 1). \quad (48)$$

What we have defined as D' in (1) above is the right member of (48) as computed from a single trial.

VI. A Third and Better Upper Bound for δ ; Further Possibilities

It is helpful in discussing the bounds to consider first the special case of scoring where $m_i \equiv 1$. For this case, (36) becomes

$$E_i \gamma_{x_{ij}x_{ig}} \leq \sqrt{\xi_g(1 - \pi_i)} \quad (m_i \equiv 1, g > j), \quad (49)$$

while (40) becomes

$$E_i \gamma_{x_{ij}x_{ig}} \leq \frac{1}{2} \sqrt{\xi_i(1 - \pi_i)} \quad (m_i \equiv 1, g > j). \quad (50)$$

Which of these provides a better bound to $E_i \gamma_{x_{ij}x_{ig}}$? That is, which has the smaller right member?

In one respect, inequality (50) is better than (49): it has the factor $1/2$.

Clearly, (49) will be better than (50) if and only if $\xi_g < \xi_i/4$ ($g > j$). Therefore, if we define ϵ_{ig} by

$$\epsilon_{ig} = \begin{cases} \sqrt{\xi_g(1 - \pi_i)} & \text{if } \xi_g < \xi_i/4 \\ \frac{1}{2}\sqrt{\xi_i(1 - \pi_i)} & \text{if } \xi_g \geq \xi_i/4 \end{cases} \quad (m_i \equiv 1, g > j), \quad (51)$$

then we have an improved bound for δ :

$$\delta \leq 2 \sum_{i=1}^n \left(\sum_{g=i+1}^n \epsilon_{ig} \right). \quad (52)$$

Inequality (52) is sharper than either (38) or (48). More generally, if $m_i \neq 1$, we can define ϵ_{ig} to be larger of the two right members of (36) and (40) and again write (52).

It is interesting that the matrix of the average error covariances, that is, of the $E \gamma_{x_i x_i g}$ is bounded in both (40) and (36) by a *simplex* matrix as defined in (4). A simplex matrix is a symmetric matrix whose elements are products of the form $a_i b_g$ ($g > j$). In (36), we can write $a_i = m_i \sqrt{1 - \pi_i}$ and $b_g = \sqrt{m_g \xi_g}$, while in (40) we can write $a_i = \sqrt{m_i \xi_i (1 - \pi_i)}$ and $b_g = \frac{1}{2} m_g$.

There are important special kinds of tests which necessarily have *internal* simplex features and not only a simplex type of upper bound matrix for error covariances. Three such are: (a) a pure speed test (where everything attempted is done correctly); (b) a test composed of a single question like "Write down all the words you can that begin with the letter 't'"; (c) a power test in which, if a person decides not to try the j th item, he will try no more items. In each such case, it follows from notation (18) that

$$p_{ijk} p_{igk} = p_{igk} \quad (g > j). \quad (53)$$

Condition (53) states that if person i does not attempt item j , then he does not attempt any items beyond j . Or if he cannot produce j words beginning with the letter t , he cannot produce g words where $g > j$.

Multiplying (53) through by x_{igk} and using (20)

$$p_{ijk} x_{igk} = x_{igk} \quad (g > j). \quad (54)$$

Using (54) in (21) yields

$$X_{ig}^{(i)} = X_{ig}/P_{ii} \quad (g > j), \quad (55)$$

showing that the inequality in (28) cannot be improved; its upper limit is actually attained in our present special case of (55). Certain further inequalities above correspondingly become equalities. Furthermore, it becomes feasible to obtain better bounds than those based on λ_3^* by pivoting instead on λ_2^* of (3).

To return to the general case where (53) does not hold, further formulas

are also possible of the split-half type, based on λ_i^* of (3). The new δ to be bounded consists of the covariance between *two sums of errors*, one sum from each half of the split. This covariance is always a *sum of item error covariances*, and can be bounded immediately by using our formulas for ϵ_{ij} above. The ϵ_{ij} should be summed the way the split calls for, and the result can be used to bound δ in the formula for λ_i^* . Using split-halves, while it requires only two sub-variances to be computed, does not avoid the problem of taking into account the possible experimental dependence between the halves, and this can be studied rigorously only itemwise, as through ϵ_{ij} .

Erratum in Guttman, Louis. Reliability formulas that do not assume experimental independence. *Psychometrika*, 1953, 18, 225-239.

On page 231 in formula (21) and in each of the preceding two lines, X_i should replace x_i throughout.

REFERENCES

1. Cronbach, L. J., and Warrington, W. G. Time-limit tests: estimating their reliability and degree of speeding. *Psychometrika*, 1951, 16, 167-187.
2. Guttman, L. A basis for analyzing test-retest reliability. *Psychometrika*, 1945, 10, 255-282.
3. Guttman, L. Reliability formulas that do not assume experimental independence. *Psychometrika*, 1953, 18, 225-239.
4. Guttman, L. A new approach to factor analysis: the radex. In Lazarsfeld *et al.*, *Mathematical thinking in the social sciences*. The Free Press, 1954.

Manuscript received 3/1/54

Revised manuscript received 8/31/54

A MATHEMATICAL MODEL FOR CONDITIONING*

G. W. BOGUSLAVSKY
CORNELL UNIVERSITY

It is postulated that occurrence of a conditioned response depends on recurrence of one of a finite number of specific vigilance reactions. Number of trial on which a conditioned response occurs is shown to be a sufficient statistic for estimating the number of such vigilance reactions. The hypothesis is tested by noting whether numbers of trials on which conditioned responses occur fall within confidence intervals determined on the basis of a selected sufficient statistic. Applications of the model to psychological research are suggested.

I. *Introduction and Postulates*

Systematic treatment of behavior has generally followed the pattern of functional relation between stimuli and responses, with intervening processes inferred from these two variables and viewed as theoretical constructs. There is reason to believe, however, that in some instances such processes may be treated as independent events. The specific reference is to behavior patterns which Pavlov grouped under the term "orienting reflex" (13, p. 134). Though Pavlov insisted that these disappear with the progress of conditioning (14, p. 94), Guthrie has presented a convincing argument to the contrary (5, p. 74), and one of Pavlov's own statements (12, p. 385) may be interpreted as refuting the original thesis.

A series of observations at the Cornell Behavior Farm has led the author to conclude that occurrence of orientation to conditioned stimuli is the rule rather than the exception. The more conspicuous features of this phenomenon are: circumscribed variability of pattern, synergy of action, facilitating effect of the general static reaction on the ensuing activity, and autonomic concomitants manifested by changes in respiration and cardiac output. All of these observations have either direct or inferential support in scientific literature (4, p. 3; 16, pp. 129, 305, 342; 2, p. 505; 19, p. 13; 7, p. 668; 10, p. 139).

It is apparent that inclusion of orienting behavior in a theoretical model would improve accuracy of prediction. However, since precise designation of the reactions and of the conditions governing their emergence is impractical, the use of monotonic functions to express relations between

*From a doctoral dissertation at Cornell University. The author wishes to acknowledge the invaluable advice and help of Professor H. S. Liddell, under whose direction this research was conducted. A special debt of gratitude is due to Dr. Jack Kiefer of the Cornell department of mathematics, whose skill and interest aided materially in the development of the mathematical portions of this paper.

variables must be abandoned. Accordingly, the present model has been designed on the basis of a system of operations which do not depend on the full knowledge of antecedent conditions.

In view of the limited connotation of the term "orienting," the author will follow Liddell's precedent (9, p. 160) in referring to the animal's immediate responses to the conditioned stimulus as *specific vigilance reactions*, or, in abbreviated form, as SVR's.

The following postulates state formally the author's theoretical position:

1. *In a given situation a supraliminal sensory stimulus evokes in an organism one of a bounded set of N discrete and mutually exclusive specific vigilance reactions.*

2. *Sensory stimulation immediately consequent upon the performance of each specific vigilance reaction becomes a conditioned stimulus for any response with which it is contiguous.*

The first proposition implies that, with the occurrence of each stimulus, the N members of the set of SVR's compete one against another, with the ultimate outcome determined by unspecified factors extraneous to the stimulus. As long as the respective probabilities of the several outcomes are unknown, the author must assume, in the present development, that the outcomes are equally likely. The proposition does not, however, preclude formulations based on empirically evaluated probabilities of identifiable intervening variables, though the problem of development along these lines would be considerably more difficult.

The second proposition implies that appearance of a conditioned response is contingent on the recurrence of any member of the set. In the present development the first recurrence is assumed to be the necessary and sufficient condition. The assumption of the *first* recurrence as the necessary and sufficient condition is based on arguments appearing in psychological literature (5). A development of the theoretical position, stated in the two postulates, on the assumption of the necessity and sufficiency of n th recurrence is no less feasible, though much more cumbersome.

The implications of the two postulates will now be examined with the aid of the classical occupancy problem serving as the model.

II. Probability Distribution of n_k

Stimuli are presented one at a time, independently of each other, the probability of a given stimulus evoking each of the SVR's being $1/N$.

An *instance of recurrence* is defined as evocation of an SVR which had been evoked on one or more preceding trials. The statement " k instances of recurrence" refers to the number of trials characterized by such recurrences. It does not imply that the same SVR has occurred k times; the number of SVR's involved in k instances of recurrence may have any integral value from 1 to k inclusive.

The variable n_k is defined to be the number of the trial on which the k th instance of recurrence takes place. By deduction from the postulates it is also the number of the trial on which the conditioned response appears for the k th time.

As an illustration of the preceding definitions, consider the four suits of a deck of cards as representing four different SVR's. In drawing with replacements the sequence H, S, H, D, S, S, C , one obtains three instances of recurrence: that of H on the third drawing, and that of S on the fifth and sixth drawings. Thus, $n_1 = 3$, $n_2 = 5$, and $n_3 = 6$. Since all suits appeared during the sequence, all subsequent drawings will be instances of recurrence.

Because n_k trials include k instances of recurrence, the total number of different SVR's evoked is $n_k - k$. Also, since the n_k th trial is one during which an instance of recurrence takes place, the total number of different SVR's evoked immediately prior to that trial is also $n_k - k$. Thus, at the end of the $(n_k - 1)$ th trial as well as at the end of the n_k th trial, the animal has in its repertory $N - n_k + k$ different unevoked SVR's. Accordingly, the probability distribution of n_k may be written as

$$P_N\{n_k = j\} = (\text{probability that after } j - 1 \text{ trials the animal's repertory contains } N - j + k \text{ different unevoked SVR's}) \times (\text{probability that on } j\text{th trial one of the previously evoked } j - k \text{ SVR's is elicited again}).$$

Evaluations of these two probabilities have been derived by Feller (3, pp. 69, 313). Allowing for differences in symbols, the substitution yields

$$\begin{aligned} P_N\{n_k = j\} &= \left[\binom{N}{N - j + k} \sum_{\nu=0}^{j-k} (-1)^\nu \binom{j-k}{\nu} \right. \\ &\quad \left. \cdot \left(1 - \frac{N - j + k + \nu}{N} \right)^{i-1} \right] \left(\frac{j-k}{N} \right) \\ &= \binom{N-1}{j-k-1} \sum_{\nu=0}^{j-k} (-1)^\nu \binom{j-k}{\nu} \left(\frac{j-k-\nu}{N} \right)^{i-1}. \end{aligned} \quad (1)$$

For the cumulative distribution of n_k another readily verifiable function derived by Feller (3, p. 77) is directly applicable.

$$\begin{aligned} P_N\{n_k \leq j\} &= \text{probability that by the end of } j\text{th trial there have been } k \text{ or more instances of recurrence of SVR's} \\ &= \text{probability that by the end of } j\text{th trial there are at least } N - j + k \text{ different SVR's not yet evoked} \\ &= \binom{N}{N - j + k} \sum_{\nu=0}^{j-k} (-1)^\nu \binom{j-k}{\nu} \\ &\quad \cdot \left(\frac{j-k-\nu}{N} \right)^i \left(\frac{N - j + k}{N - j + k + \nu} \right). \end{aligned} \quad (2)$$

III. Sufficiency of the Statistic n_k

For a discrete case, $t_n = t(x_1, x_2, \dots, x_n)$ is said to be a sufficient statistic for the parameter N if, whenever $p_N(t_n) > 0$, the conditional probability function $p_N(x_1, \dots, x_n | t_n)$ does not depend on N .

A necessary and sufficient condition that t_n be sufficient is that the joint probability function of x_i 's can be written

$$p_N(x_1, \dots, x_n) = g(t_n, N) \cdot h(x_1, \dots, x_n), \quad (3)$$

where $g \geq 0$, $h \geq 0$, g depends on x_i 's only through the function t_n , and h depends on x_i 's in any way, but does not depend on N .

It will now be shown that the joint probability distribution of the variables n_i , $1 \leq i \leq k$, can be written in the form (3), where n_k takes the place of t_n , and that, therefore, n_k is a sufficient statistic for the parameter N .

Let $X_i + 1$ be the number of trials between $(i - 1)$ th and i th instances of recurrence of SVR's (not counting the former, but counting the latter). For $i = 1$,

$P_N\{X_1 = x_1\}$ = probability that the first x_1 stimuli evoke x_1 different SVR's, and the $(x_1 + 1)$ th stimulus evokes an SVR which had occurred earlier,

$$\begin{aligned} &= 1 \cdot \frac{N-1}{N} \cdot \frac{N-2}{N} \cdot \dots \cdot \frac{N-x_1+1}{N} \cdot \frac{x_1}{N} \\ &= \frac{N!x_1}{N^{x_1+1}(N-x_1)!} \end{aligned} \quad (4)$$

The conditional probability distributions for the variables X_i , $i > 1$, may be derived in an analogous manner. For $i = 2$,

$$\begin{aligned} p_N(x_2 | x_1) &= \frac{N-x_1}{N} \cdot \frac{N-x_1-1}{N} \cdot \dots \cdot \frac{N-x_1-(x_2-1)}{N} \cdot \frac{x_1+x_2}{N} \\ &= \frac{(N-x_1)!(x_1+x_2)}{N^{x_2+1}(N-x_1-x_2)!} \end{aligned} \quad (5)$$

Designating for convenience

$$S_v = \sum_{i=1}^v x_i, \quad (6)$$

the conditional probability distribution for the general case is

$$\begin{aligned} p_N(x_v | x_1, \dots, x_{v-1}) &= \left(\prod_{i=0}^{x_v-1} \frac{N-S_{v-1}-i}{N} \right) \left(\frac{S_v}{N} \right) \\ &= \frac{(N-S_{v-1})!S_v}{N^{x_v+1}(N-S_v)!} \end{aligned} \quad (7)$$

The joint probability distribution of the variables X_i , $1 \leq i \leq k$, is obtained by multiplying the probability distribution of the initial variable X_1 by the product of the conditional probability distributions of the remaining variables X_i , $i > 1$,

$$\begin{aligned} p_N(x_1, \dots, x_k) &= p_N(x_1) \cdot \prod_{i=2}^k p_N(x_i | x_1, \dots, x_{i-1}) \\ &= \frac{N!x_1}{N^{x_1+1}(N-x_1)!} \cdot \prod_{i=2}^k \frac{(N-S_{i-1})!S_i}{N^{x_i+1}(N-S_i)!} \\ &= \frac{N!}{N^{S_k+k}(N-S_k)!} \cdot \prod_{i=1}^k S_i \end{aligned} \quad (8)$$

$$= \frac{N!}{N^{n_k}(N-n_k+k)!} \cdot \prod_{i=1}^k (n_i - i), \quad (9)$$

where $n_i = S_i + i$, the relationship being derived from definitions presented earlier.

Inspection of (9) shows that the function represented by the first factor does not depend on the variables n_i , $i < k$, except through n_k ; the function represented by the remaining product does not depend on N . Thus the criterion of (3) is satisfied, showing that n_k is a sufficient statistic. This implies that estimation of the parameter N may be made solely from the knowledge of n_k , and that knowledge of the values of n_i , $i < k$, provides no additional information.

IV. Tests of Hypotheses

The foregoing discussion indicates that description of the progress of conditioning is reducible to the single parameter N . Although many specific vigilance reactions may be identified with accuracy from their unique postural components, precise evaluation of N by direct observation is hardly feasible at this stage. Accordingly, the experimenter must resort to estimation of N from information gathered during the process of conditioning. One approach is to proceed with presentation of stimuli until the k th occurrence of the conditioned response, k being a previously selected constant. With an observed value j of n_k , either point estimation or interval estimation of the parameter N may be made. One possible procedure for point estimation, the so-called maximum-likelihood procedure, involves the use of (1), where N is chosen so that the value of $P_N\{n_k = j\}$ is maximized. Procedure for interval estimation will be described in the section on confidence intervals.

Another approach to the evaluation of N consists of extending the experiment to the stage at which the animal performs a conditioned response unfaithfully on every trial. From the postulates it follows that those trials on which no conditioned responses occur are trials characterized by novel

SVR's. Hence, the total number of trials on which no conditioned responses occurred is directly equivalent to N . The practical problem in this approach is clearly that of defining the criterion of conditioning. The following discussion deals with tests of hypotheses concerning the parameter N . While the subject is one of intrinsic interest, the discussion is introduced at this point as a preliminary step in the development of a procedure for interval estimation.

Let $N_0 < N_1$ be two specified positive integers, and designate by H_0 the hypothesis that $N = N_0$ and by H_1 the alternative hypothesis that $N = N_1$. A test of $H_0 : N = N_0$ against $H_1 : N = N_1$ involves a single alternative and a single observation. Accordingly, construction of such a test consists of selecting a critical region such that

$$\frac{P_{N_1}\{n_k = j\}}{P_{N_0}\{n_k = j\}} > c, \quad (10)$$

where c is chosen so that the probability of the critical region under H_0 is θ .

Substitution from (1) into (10), with the factors not involving N canceling out, yields

$$\frac{N_0^j \cdot N_1! \cdot (N_0 - j + k)!}{N_1^j \cdot N_0! \cdot (N_1 - j + k)!} = L(j). \quad (11)$$

It will be noted that

$$\frac{L(j+1)}{L(j)} = \frac{N_0}{N_1} \cdot \frac{N_1 - j + k}{N_0 - j + k} > 1, \quad (12)$$

the last inequality being a consequence of the fact that $N_1 > N_0$ and $j > k$. Hence $L(j)$ is a monotonically increasing function of j , which implies that the most powerful critical region is one for which n_k exceeds a selected constant. This is a uniformly most powerful test, for the specified probability of Type I error, of the hypothesis $H_0 : N = N_0$ against $H_1 : N = N_1$, since the constant depends only on N_0 . It is also a uniformly most powerful test of the hypothesis $H_0 : N \leq N_0$ against $H_1 : N > N_0$, since the probability of Type I error for any $N < N_0$ is less than that for N_0 under this test.

The criterion stated in (10) is thus equivalent to the rule to reject $H_0 : N = N_0$ if n_k takes on a value j such that

$$j \geq b, \quad (13)$$

where b is chosen so that

$$P_{N_0}\{n_k \geq b\} = \theta. \quad (14)$$

Similarly, a uniformly most powerful test of the hypothesis $H_0 : N = N_0$ (or $N \geq N_0$) against $H_1 : N < N_0$ is given by the rule to reject $H_0 : N = N_0$ if n_k takes on a value j such that

$$j \leq a, \quad (15)$$

where a is chosen so that

$$P_{N_0}\{n_k \leq a\} = \theta. \quad (16)$$

For testing the hypothesis $H_0: N = N_0$ against $H_1: N \neq N_0$, a uniformly most powerful test of size θ does not exist. An approximation to the best unbiased test is to choose two numbers a and b such that

$$P_{N_0}\{n_k \leq a\} = P_{N_0}\{n_k \geq b\} = \frac{\theta}{2}, \quad (17)$$

and to reject $H_0: N = N_0$ by the criteria stated in (13) and (15).

V. Confidence Intervals

The best one-sided confidence limits on N can be easily constructed. Thus, the rule of (13) and (14) is equivalent to the rule of accepting $H_0: N = N_0$ if n_k takes on a value j such that

$$j \leq b - 1, \quad (18)$$

where b is chosen so that

$$P_{N_0}\{n_k \leq b - 1\} = 1 - \theta. \quad (19)$$

With θ selected arbitrarily, values of b are calculated for different values N_0 by means of (2), where j is the symbol for $b - 1$. For each N_0 let $b(N_0)$ designate the value thus calculated. Each value $b(N_0) - 1$ is now plotted as the ordinate above the corresponding value N_0 of N on the horizontal axis. Then, for each value N_0 of N ,

$$P_{N_0}\{n_k \leq b(N_0) - 1\} = 1 - \theta. \quad (20)$$

Let $L(n_k)$ be the inverse of $b(N_0) - 1$, thus designating N as a function of n_k . Since $b(N_0) - 1$ is the maximal value corresponding to N_0 , given the condition (19), clearly, for a specified n_k , $L(n_k)$ is the minimal value; i.e., for the selected θ , the value $b(N_0) - 1$ of n_k may be obtained only with those values of N which are equal to or exceed $L(n_k)$. Thus, (20) is equivalent to

$$P_{N_0}\{N_0 \geq L(n_k)\} = 1 - \theta. \quad (21)$$

The last equation implies that, whatever the true value N_0 of N , the probability is $1 - \theta$ that the chance variable n_k will come up so that $N_0 \geq L(n_k)$. In other words, $L(n_k)$ is a one-sided confidence limit on N of confidence coefficient $1 - \theta$.

Similarly, for each value N_0 of N a value $a(N_0)$, corresponding to the

a of inequality (15), is calculated. The values $a(N_0) + 1$, plotted as before, yield the relation

$$P_{N_0}\{n_k \geq a(N_0) + 1\} = 1 - \theta, \quad (22)$$

and its inverse form

$$P_{N_0}\{N_0 \leq U(n_k)\} = 1 - \theta, \quad (23)$$

the last equation implying that n_k will come up so that the probability of $U(n_k)$ being equal to or greater than the true value of N is always $1 - \theta$, whatever be that true value of N .

The methods for constructing one-sided confidence limits are best because the corresponding tests of hypotheses from which they are derived are uniformly most powerful. Since there is no uniformly most powerful test of $H_0 : N = N_0$ against $H_1 : N \neq N_0$, the two-sided confidence interval may be constructed by a procedure approximating an unbiased test, or one with the shortest acceptance region. For each value N_0 of N a value $a(N_0)$ and a value $b(N_0)$, corresponding to a and b of (17), are calculated. These are such that, whatever the true N_0 ,

$$P_{N_0}\{a(N_0) + 1 \leq n_k \leq b(N_0) - 1\} = 1 - \theta, \quad (24)$$

and, inversely,

$$P_{N_0}\{L(n_k) \leq N_0 \leq U(n_k)\} = 1 - \theta. \quad (25)$$

Since n_k is a discrete chance variable, it is not always possible to find values of a and b which yield exactly θ for each N_0 . A conservative procedure of designating the limiting values $L(n_k)$ and $U(n_k)$ of (25) would be to state the integral values of N which lie nearest the limits, outside the interval defined by n_k and $1 - \theta$.

Figure 1 gives values of j_U and j_L for $1 \leq k \leq 5$, such that

$$P_{N_0}\{n_k \leq j_U\} = P_{N_0}\{n_k \geq j_L\} \doteq .90, \quad (26)$$

with j_U and j_L chosen so as to make this probability as little greater than .90 as possible. The upper limits are designated U and the lower L . The number preceding U or L refers to the value of k for which the limit was computed. Thus,

$$P_{200}\{n_2 \leq j_U\} \doteq .90 \quad (27)$$

is interpreted as "the probability is .90 that, given $N = 200$, the second recurrence of an SVR will take place no later than j_U th trial."

Locating 200 on the horizontal axis, one proceeds vertically to the curve $2U$, and thence horizontally to the vertical axis which is met at $n_k = 39$; the latter is the value which satisfies the condition (27). Similarly, the lower limit of n_2 may be evaluated, showing that

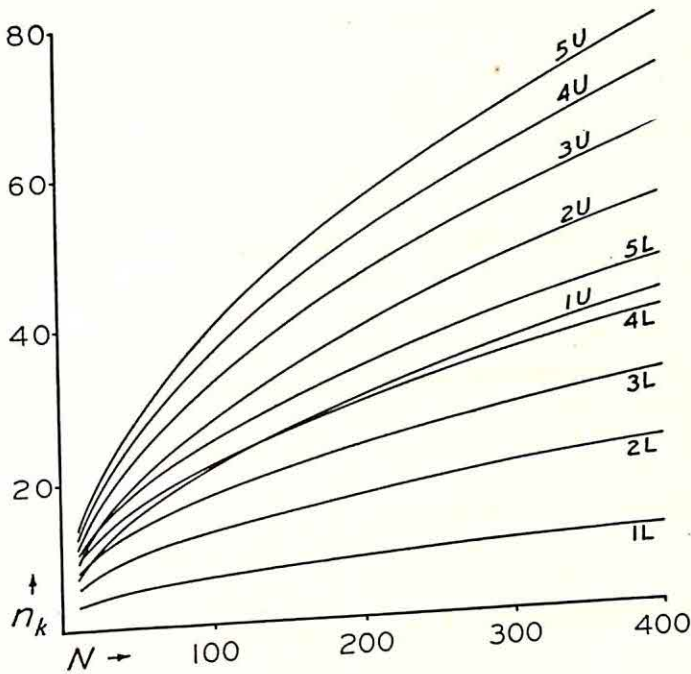


FIGURE 1
Confidence Limits on N

$$P_{200}\{n_2 \geq 17\} \doteq .90. \quad (28)$$

Combining (27) and (28), one obtains an approximation to be used in constructing the best unbiased confidence interval on N , defined by the parameter 200 and the confidence coefficient .80. Thus,

$$P_{200}\{17 \leq n_2 \leq 39\} \doteq .80. \quad (29)$$

The same procedure is followed in locating confidence intervals for other values of N and k within the limits of the chart.

In a typical conditioning situation N is unknown, and the problem is one of estimating this parameter from the variates n_k . The chart shown in Figure 1 fulfills this function if used inversely. Assuming, for example, that the fifth conditioned response occurs on trial 20, a horizontal line is drawn from the vertical axis at 20, and ordinates are dropped from its intersections with $5U$ and $5L$. These meet the horizontal axis at 25 and 60. The estimated interval on N is now stated as follows: "Because the fifth conditioned response occurred on trial 20, it may be stated with 80% confidence that the interval 25 to 60 includes the true value of the total number of specific vigilance reactions possessed by the organism in the given situation."

VI. A Test of the Model

The model may be tested by observing whether, for a given N , the joint distribution of n_i 's is within the region of acceptance selected in such a manner that its total probability is $1 - \theta$. While this method of testing is highly desirable, it is not feasible at the present stage of the development of the model, inasmuch as it depends on the knowledge of the true value of N . It has been possible, however, to construct a test which does not depend on the knowledge of this parameter. The procedure for such a test is described below.

In Section III it was shown that, if the model is true, the statistic n_k is sufficient, and that, consequently, the conditional probability distribution

$$P_N\{n_1 = j_1, n_2 = j_2, \dots, n_{k-1} = j_{k-1} \mid n_k = j_k\}$$

does not depend on N .

This distribution is given by

$$\frac{p_N(j_1, j_2, \dots, j_k)}{p_N(j_k)} = \frac{\prod_{v=1}^{k-1} S_v}{\sum_{S_v} \prod_{v=1}^{k-1} S_v}, \quad (30)$$

where the expression on the right is derived from (8), and the summation in the denominator is taken over all possible sets of values of S_v such that

$$1 \leq S_1 \leq \dots \leq S_v \leq \dots \leq S_{k-1} \leq j_k - k. \quad (31)$$

An appropriate summation of the numerator in (30) yields the conditional probability distribution of a single variable n_i . Thus,

$$p(j_i \mid j_k) = \frac{S_i \left(\sum_{S_v} \prod_{v=1}^{i-1} S_v \right) \left(\sum_{S_v} \prod_{v=i+1}^{k-1} S_v \right)}{\sum_{S_v} \prod_{v=1}^{k-1} S_v}, \quad (32)$$

where sums in the numerator are taken over all possible sets of values of S_v such that, in the summation of the first product,

$$1 \leq S_1 \leq \dots \leq S_i, \quad (33)$$

and, in the summation of the second product,

$$S_i \leq S_{i+1} \leq \dots \leq S_{k-1} \leq j_k - k. \quad (34)$$

Finally, a summation of (32) over the indicated values of the variable in the numerator yields the cumulative conditional probability distribution of n_i ,

$$P\{n_i \leq j_i \mid n_k = j_k\} = \frac{\sum_{s_i=1}^{j_i-1} \left[S_i \left(\sum_{s_r} \prod_{r=1}^{i-1} S_r \right) \left(\sum_{s_r} \prod_{r=i+1}^{k-1} S_r \right) \right]}{\sum_{s_r} \prod_{r=1}^{k-1} S_r}, \quad (35)$$

with conditions for the other sums defined by (31), (33), and (34).

Equation (35) provides a direct test of the model. Thus, if the model is true, the observed value x of n_i will be such that, with probability $1 - \theta$,

$$\frac{\theta}{2} \leq P\{n_i \leq x \mid n_k = j_k\} \leq 1 - \frac{\theta}{2}. \quad (36)$$

Selecting $k = 5$, and designating the critical area $\theta = .20$, test observations were made on four goats in a conditioning situation. An auditory signal served as the neutral stimulus, and flexion of the right foreleg, unconditionally evoked by an electric shock, as the response. The results are presented in Table 1.

TABLE 1
Ordinal Numbers of Trials on Which CR's Occurred
and Their Cumulative Conditional Probabilities

i	H		L		Y		F	
	$n_5 = 16$		$n_5 = 17$		$n_5 = 20$		$n_5 = 30$	
	j_i	P	j_i	P	j_i	P	j_i	P
1	9	.90	9	.84	8	.57	21	.97
1	9	.90	9	.84	14	.80	22	.83
2	10	.65	12	.85	16	.70	23	.53
3	11	.35	14	.82	18	.60	28	.73
4	15	1.00	15	.54				

Capital letters in the top row of Table 1 are code letters of the four animals. Beneath each code letter is the number of the trial on which that animal gave the fifth CR. The column labeled i contains ordinal numbers of the first four CR's. Columns labeled j_i give the numbers of trials on which i th CR's occurred. Columns labeled P give the cumulative conditional probabilities calculated by means of (35). Thus, by way of illustration, goat Y gave its fifth CR on trial 20, and its third CR on trial 16. The probability that an animal which gave its fifth CR on trial 20 should have given its third CR no later than trial 16 is .70, which is well within the arbitrarily selected acceptance interval of .10 to .90.

Inspection of Table 1 shows that only two of the sixteen probabilities fail to meet the criterion of acceptance. These are probabilities computed for the fourth CR of H and the first CR of F . The extreme value of H may, however, be explained by the fact that the conditional probability of the fourth CR having occurred exactly on trial 15, given that the fifth CR took place on trial 16, is .49. Of course, the four tests corresponding to different

rows in any column of Table 1 are not independent of each other; however, as a rough indication of the validity of the model, Table 1 presents a convincing demonstration.

At this point it may be noted that selecting regions of acceptance for each n_i separately rather than designing a single region for their joint distribution actually increases the size of the test; i.e., the total probability that at least one of the rows in any given column would give a result leading to rejection of the model when it is true is greater than θ . The effect is illustrated in Figure 2, where the test is applied to the case of n_1 and n_2 , given n_k , $k > 2$.

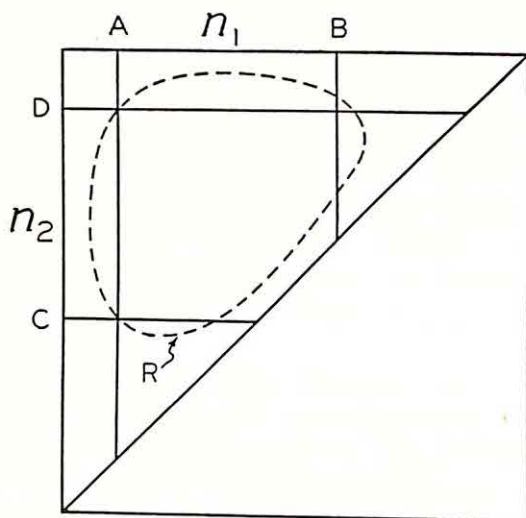


FIGURE 2

Schematic Representation of Regions of Acceptance

Since n_2 is greater than n_1 , the joint probabilities are greater than zero above the diagonal, and zero elsewhere. R represents the region of acceptance in which the sum of joint probabilities is $1 - \theta$. Construction of R involves laborious mathematical computations which would be hardly justifiable in the absence of specified alternatives to the model. Since a test of the model at this stage is intended merely as a detecting device of obvious fallacies, if any, the substitute procedure, defined by (35), consists of selecting separate regions of acceptance for n_1 and n_2 , such that in each case the sum of marginal probabilities excluded on each side of the region is $\theta/2$. This construction of the regions of acceptance yields probability intervals which are approximately the shortest possible, since the ordinates which form the limits of this region are approximately equal. Such a test should give good power against most reasonable alternatives to the model.

In Figure 2 the area bounded by the vertical lines A and B is the region of acceptance for n_1 , and the sum of joint probabilities within this area is

$1 - \theta$. Similarly, the area bounded by the horizontal lines C and D is the region of acceptance for n_2 , and the sum of joint probabilities within this area is also $1 - \theta$. With this construction the condition for non-rejection is that both n_1 and n_2 occur within the rectangle bounded by the four lines. It is clear, however, that the sum of joint probabilities within this rectangle is less than $1 - \theta$. Hence, the size of the test is actually greater than θ ; or, with a diminished probability of Type II error, the power of the test is somewhat higher than that of the precise test based on the region of acceptance R .

VII. *Conclusions*

The mathematical model for conditioning has been presented primarily as an illustration of a technique for the treatment of intervening variables, rather than as a substitute for the existing systems. Since the model does not demand rigor in the definition of these variables, it has possibilities of adaptation to theories of behavior which regard autonomous central processes as crucial. The author is currently engaged in one such adaptation, extending the model to include problems in discrimination learning. The extension should furnish a method for treating Krechevsky's "hypotheses" (8) as stochastic variables, thus providing a testable alternative to Spence's model (18).

In another application, the model provides, in the value of N , a measure for the study of individual differences. Organisms requiring many trials to reach a stipulated criterion will, on the average, yield larger estimates of N . Since N is the sole parameter involved, the phenomenon of slow learning may be interpreted as inability on the part of the organism to restrict its range of vigilance to the situation at hand. Since, however, the author knows of no way to test the latter inference, it must remain, at least for the present, on the level of intuitive generalization. On the other hand, estimates of N , based on a limited number of trials, are intended to provide uniform quantitative indices for a variety of psychological investigations, ranging from selection of stratified samples to problems of heredity and environment. Conversely, for the same organism, the parameter may furnish a comparative estimate of the efficacy of various learning situations.

Studies involving large populations are often prohibitive because of time and effort required to train each subject to a criterion of mastery. Instead, in training each subject to a predetermined number of conditioned responses, one is able to make a reasonably accurate quantitative estimate of the subject's susceptibility to training, with a substantial economy in labor. Furthermore, the potentialities of the parameter for the construction of gradients of similarity may lead eventually to a reexamination of the phenomena of generalization and pseudo-conditioning in the light of mediating factors susceptible to systematic treatment.

The model may be regarded from one of two contrasting theoretical

positions. One may either take the stand advocated by Skinner (17) and view the parameter N purely as "a formal representation of the data reduced to a minimal number of terms," or one may follow the course suggested by Pratt (15) and postulate independent existence of neurophysiological events corresponding to this parameter. The author leans towards the latter point of view because of its potentiality as a source of future hypotheses. Moreover, improved techniques of observing and recording specific vigilance reactions may ultimately lead to an independent estimate of the parameter N , thus serving as the second of "at least two methods" stipulated by Bridgman (1) "of getting to the terminus."

REFERENCES

1. Bridgman, P. W. Some general principles of operational analysis. *Psychol. Rev.*, 1945, 52, 246-249.
2. Denny-Brown, D. (Ed.) Selected writings of Sir Charles Sherrington. New York: Hoeber, 1940.
3. Feller, W. An introduction to probability theory and its applications. New York: Wiley, 1950, Vol. I.
4. Fraenkel, G. S., and Gunn, D. L. The orientation of animals. Oxford: Clarendon Press, 1940.
5. Guthrie, E. R. The psychology of learning. New York: Harper, 1952.
6. Halmos, P. R., and Savage, L. J. Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *Ann. math. Statist.*, 1949, 20, 225-241.
7. Hoff, H. E. Cardiac output: regulation and estimation. In J. F. Fulton (Ed.), A textbook of physiology. Philadelphia: Saunders, 1950. Pp. 660-680.
8. Krechevsky, I. "Hypotheses" in rats. *Psychol. Rev.*, 1932, 39, 516-532.
9. Liddell, H. S. Some specific factors that modify tolerance for environmental stress. *Proc. Ass. Res. nerv. ment. Dis.*, 1949, 29, 155-171.
10. Liddell, H. S. The influence of experimental neuroses on respiratory function. In H. A. Abramson (Ed.), Somatic and psychiatric treatment of asthma. Baltimore: Williams and Wilkins, 1951. Ch. 6.
11. Mood, A. M. Introduction to the theory of statistics. New York: McGraw-Hill, 1950.
12. Pavlov, I. P. Conditioned reflexes. London: Oxford Univ. Press, 1927.
13. Pavlov, I. P. Lectures on conditioned reflexes. New York: International, 1928.
14. Pavlov, I. P. The reply of a physiologist to psychologists. *Psychol. Rev.*, 1932, 39, 91-127.
15. Pratt, C. C. Operationism in psychology. *Psychol. Rev.*, 1945, 52, 262-269.
16. Sherrington, C. S. The integrative action of the nervous system. New Haven: Yale Univ. Press, 1947.
17. Skinner, B. F. Are theories of learning necessary? *Psychol. Rev.*, 1950, 57, 193-216.
18. Spence, K. W. The nature of discrimination learning in animals. *Psychol. Rev.*, 1936, 43, 427-449.
19. Voitonis, N. Predistoriia intellekta. Moscow: Akad. Nauk, 1949.

Manuscript received 2/9/54

Revised manuscript received 7/30/54

TWO MODELS OF GROUP BEHAVIOR IN THE SOLUTION OF EUREKA-TYPE PROBLEMS*

IRVING LORGE

AND

HERBERT SOLOMON

TEACHERS COLLEGE, COLUMBIA UNIVERSITY

A study by Shaw (7) some twenty years ago is frequently cited by social scientists to support the generalization that groups are superior to individuals in problem-solving. Shaw suggests that personal interaction within the group is responsible for the superior performance of groups. This article re-examines her data in the light of two models which propose that the difference in quality of solution between group and individual performance is solely a matter of ability. It is shown that Shaw's data may be considered to have been an outcome of behavior postulated by the models. Since Shaw's observations relate to a special population and to special kinds of problems, the proposed models may not be appropriate under differing experimental conditions. In fact, Lorge *et al.* (4) have indicated that experimental demonstration of the superiority of groups over individuals in problem-solving depends not only on the kind of group but also on the kind of problem to be solved. In addition, the diversity of transfer of training for groups and for individuals is considered.

Introduction

Since this article treats only the data from the first half of the Shaw experiments, a brief description of this part will be given. Three problems (3), each a well-known mathematical puzzle involving the transport of objects under certain constraints, were given to groups and to individuals. The first, known historically as the Tartaglia, requires the transport of three jealous husbands and their three beautiful wives across a river in a boat holding just three at a time, under the constraint that no husband will allow his wife in the presence of another man unless he is also present, and with the specification that only husbands can row. The second problem, the historical Alcuin, is similar in that it requires the transport of three missionaries and three cannibals in a boat carrying two at a time under the constraint that missionaries may never be outnumbered by cannibals, and with the specification that all missionaries and just one cannibal have mastered the art of rowing. The third problem, the historical Tower of Hanoi, or disc problem, is similar to the previous two in that it requires the transport of three graduated discs, stacked in order of size, to another position via an intermediate way station, under the constraint that a larger disc may never

*Supported in part by the Office of Naval Research under Contract N6 onr 266 (21) and the Air Force Personnel and Training Research Center under Contract AF 18(600)-341.

be placed on a smaller one, with the specification that only one disc may be moved at a time.

Shaw's subjects were students in a social psychology class which had been divided into halves: one half being formed at random into *ad hoc* like-sex, four-member groups, and the other half serving as individuals, i.e., as controls. Thus, the performances of five groups were contrasted with those of twenty-one individuals. Each group and each individual was asked to solve all three problems in the same sequence.

A criterion for comparing group and individual performance is the contrast between the proportion of individuals and the proportion of groups successful in the solution of each problem. For Shaw's three problems, the proportions of individuals and groups mastering each solution are given in Table 1 (Columns 1 and 3). When, for each problem separately, the difference between proportions of success in groups and in individuals is tested, using an upper one-sided .05 critical region, the data for Problems I and II support the generalization of group superiority, but the difference between groups and individuals for Problem III is not statistically significant. The statistical test (2, 6) of the hypothesis that two proportions are equal is

$$z = \frac{\theta_G - \theta_I}{\sqrt{\frac{1}{N_I} + \frac{1}{N_G}}}, \quad (1)$$

where $\theta = 2 \arcsin \sqrt{p}$, p = proportion of success, N = sample size, and the subscripts I and G refer to individuals and to groups, respectively. The function z is approximately normally distributed with zero mean and unit variance under the hypothesis tested. The results of this analysis could be used to support Shaw's conclusion (7, p. 504): "Groups seem assured of a much larger proportion of correct solutions than individuals do."

Of the five groups, however, two solve *none* of the problems and two solve *all* problems. Of the twenty-one individuals, none solves more than *one* of the three problems. The fact that some groups solved none and some groups solved all the problems suggests the hypothesis that the observed group superiority is due to the abilities of the members of the group rather than personal interaction. Such an hypothesis may be expressed in terms of two ability models: (A) group superiority is a function only of the ability of one or more of its members to solve the problem without taking account of the interpersonal rejection and acceptance of suggestions among its members; (B) group superiority is a function only of the pooled abilities of its members. The latter model, B , implies that any problem may be composed of, and solved in, two or more stages. Model B , of course, reduces to Model A for one-stage problems.

Model A

Under Model A the probability of a group solution is the probability that the group contains one or more members who can solve the problem. This non-interactive ability model for any specific problem can be expressed mathematically as follows: Let

P_G = the probability that a group of size k solve the problem;
 P_I = the probability that an individual solve the problem.

Then

$$P_G = 1 - (1 - P_I)^k, \quad (2)$$

where P_G and P_I are population parameters considered fixed for the specific problem and the specific population.

Confidence in the tenability of this non-interactive ability model can be decided by testing it on the basis of sample observations. Assume N_G observations of group performance and N_I of individual performance. Then sample estimates p_G and p_I may be obtained, where p_G and p_I are the ratios of the observed successes to attempts for groups and for individuals, respectively; p_G should be compared with p_{GA} (or equivalently, p_I with p_{IA}), where

$$p_{GA} = 1 - (1 - p_I)^k, \quad (3)$$

or equivalently

$$p_{IA} = 1 - (1 - p_G)^{1/k}. \quad (3a)$$

The observed difference $(p_G - p_{GA})$ certainly can be used as a test of the model, for the smaller the observed difference, the more tenable is the model and, the larger the observed difference, the less tenable it is. If an α level of significance is used, then the model would be rejected if

$$\Pr \{(p_G - p_{GA}) > O_d\} \leq \alpha$$

and accepted otherwise, where O_d is the observed difference. A one-sided test is used since *negative* personal interaction (an unable majority preventing an able minority from solving the problem) is not anticipated in the Shaw groups, and thus the test is made most powerful against all alternatives indicating *positive* personal interaction. That is, if positive interaction does exist, the probability of rejecting Model A is higher than the probability given by a two-sided test of the same size. A similar argument holds for $(p_{IA} - p_I)$, since it is an equivalent test.

To test the existence of the model, the distribution of $(p_G - p_{GA})$ must be obtained. Although p_G and p_{GA} are independently distributed proportions, the distribution of their difference is no longer related to the standard distri-

bution of the difference of two binomials since p_{GA} is not a binomial; p_{GA} is a function of p_I , which is a binomial. This complicates obtaining the exact distribution of $(p_G - p_{GA})$ either in closed form or in a form such that existing tables may be used. Since sample sizes are small, however, it is not too tedious to compute the exact probabilities of all differences larger than the observed difference under the assumptions that (1) the model holds and (2) the nuisance parameter (either P_G or P_I) is replaced by a sample estimate.

It is interesting to note that

$$E(p_{GA}) = 1 - (1 - P_I)^4 - \frac{6P_I(1 - P_I)^3}{N_I} + \frac{P_I(1 - P_I)^2(4 - 11P_I)}{N_I^2} - \frac{P_I(1 - P_I)(1 - 6P_I + 6P_I^2)}{N_I^3},$$

and

$$\sigma_{p_{GA}}^2 = \frac{P_I(1 - P_I)}{N_I} [16(1 - P_I)^6] + \frac{f_1(P_I)}{N_I^2} + \cdots + \frac{f_6(P_I)}{N_I^7},$$

where $f_i(P_I)$, $j = 1, 2, \dots, 6$, are eighth-degree polynomials in P_I . Thus, for large N_I , p_{GA} is an unbiased estimate of P_G and its variance is $16(1 - P_I)^6 \sigma_{p_I}^2$.

For the three Shaw problems, there are six possible values for p_G and twenty-two possible values of p_I . In Problem I, for instance, the observed difference $(p_G - p_{GA})$ is .14, where p_{GA} is computed from formula (2) using the value of p_I reported by Shaw. It is necessary, therefore, to tabulate all possible differences greater than the value .14. For these tabulated differences, the probability of each is computed under the specified assumptions. The probability for each difference is the product of the probabilities that the p_G and p_{GA} involved in the difference do occur when the two assumptions hold. The probability that a p_{GA} occurs is equal to the probability that its corresponding p_I occurs. The probability for p_G and p_{GA} may be obtained readily by reference to a binomial table (5). The sum of these products of probabilities is the exact probability that an observed difference will exceed .14. In Table 1, column five gives the exact probability, P , that the observed difference $(p_G - p_{GA})$ will be exceeded by chance.

An approximation to the exact probability can be made when p_I is small enough so that p_{GA} can be approximated by kp_I , for then

$$\frac{1}{k} (2 \arcsin \sqrt{kp_I}) \quad \text{and} \quad \frac{1}{k} (2 \arcsin \sqrt{p_G})$$

are approximately normally distributed with variances

$$\frac{1}{N_I} \quad \text{and} \quad \frac{1}{k^2 N_G}, \text{ respectively.}$$

Thus, if Model A holds,

$$z = \frac{2 \arcsin \sqrt{p_G} - 2 \arcsin \sqrt{kp_I}}{\sqrt{\frac{1}{N_G} + \frac{k^2}{N_I}}} \quad (4)$$

is approximately normally distributed with zero mean and unit variance. Some liberties have been taken in this approximation by assuming kp_I to be binomial since it can assume values greater than one. This assumption, apparently, does not impair its usefulness for the Shaw experiments. In Table 1, column six gives $P' = P_r\{z > z_0\}$, where z_0 is the specific value for z corresponding to the observed difference. Notice that the approximation obviously gets better as p_I decreases.

The hypothesized non-interactional ability Model A, thus, is rejected for Problem II, but accepted as tenable for Problems I and III. For each of the three problems, however, p_G exceeds p_{GA} , suggesting that Model A might be modified and improved.

TABLE 1

	p_I	p_{IA}	p_G	p_{GA}	P	P'
Problem I	3/21 = .14	.20	3/5 = .60	.46	.38	.48
Problem II	0/21 = .00	.20	3/5 = .60	.00	.029	.023
Problem III	2/21 = .095	.12	2/5 = .40	.33	.43	.48

p_I = ratio of individual solutions to attempts

p_G = ratio of group solutions to attempts

p_{IA} = estimate of P_I from Model A and observation p_G

p_{GA} = estimate of P_G from Model A and observation p_I

P = probability ($p_G - p_{GA}$) is exceeded by chance under Model A and P_G or P_I is replaced by sample estimate

P' = approximation of P replacing p_{GA} by kp_I

Stage-wise Solutions

Within the framework of strict ability models, a modification of Model A may be made. Solution of eureka-type problems may be considered the consequence of pooling success at each of several stages of the problem. Shaw's study, indeed, suggests the plausibility of such a stage-wise model. In reporting about the erroneous moves made by her subjects in solving Problem I she states that 13 different individuals made an error in the first move, four made an error in the third move, and one made an error in the fifth. For groups, however, she reports "No group erred on the first move; one erred on the third and one on the fourth."

Shaw's description of the errors in Problem I suggests the importance of the first move, since 13 of the 21 individuals failed to make the correct first move. Each group, however, apparently had in it at least one member

who made the first move successfully since none of the five groups erred on it. Once the first move is accomplished, the difficulty of the problem changes. Five individuals who made the first move correctly did fail at subsequent stages, i.e., made the first correct move but failed at later moves. Two groups failed at some later move, suggesting that the group lacked at least one member who could accomplish some later move.

Assuming that a problem is solved in s independent stages, (not the moves Shaw mentions, since such moves may be interrelated) and assuming that Model A (equation 2) applies at each stage j , then,

$$P_G = \prod_{i=1}^s [1 - (1 - P_{I_i})^k], \quad P_I = \prod_{i=1}^s P_{I_i}, \quad (5)$$

where s is the number of stages, and P_{I_i} is the probability of success for an individual at stage j . Now for the purpose of estimating s from the Shaw data, consider the assumption that P_{I_i} is the same for each stage; thus $P_{I_i} = P_I^{1/s}$, then

$$P_G = [1 - (1 - P_I^{1/s})^k]^s. \quad (5a)$$

This assumption may possibly be unrealistic, but it is necessary to provide an estimate of s from Shaw's data.

Substituting the estimates for P_G and P_I from Shaw's Problems I and III, $s = 2$ (to the nearest integer) for both problems; for Problem I, $s = 1.6$; for Problem III, $s = 1.5$. Since the observed proportions of individual solutions for Problem II is zero, s is indeterminate. (If for Problem II, P_I is replaced by $p_{I_A} = .2$, then s is very close to 1.)

It is not too difficult to rationalize the two-stage nature of the problems. For example, in the problem of the jealous husbands and their wives, the basic first stage requires the recognition that the boat, which may carry three, must be limited to taking just a husband and his wife across the river. Once this first stage is solved, the second and final stage is analogous to repetitious knitting. It is interesting to note that if it is assumed that $p_I = .05$ and $p_G = .95$ (an indication of overwhelming group superiority through positive personal interaction among its members) then by (5a) $s = 10$ to the nearest integer, an estimate even larger than the number of moves required in some of the Shaw problems. While all possible pairs of the values p_G and p_I have not been considered, an excessively large difference gives a value of s inconsistent with a psychological analysis of the problem into steps or stages.

Model B

On both a probabilistic and a content basis, a two-stage problem may be reasonably inferred; assume now that Problems I, II, III are two-stage problems. For this situation, the population of individuals may be classified in the following way:

<i>Population Type</i>	<i>Ability</i>	<i>Proportion in the Population</i>
X_1	Solve both stages	P_1
X_2	Solve stage 1, not stage 2	P_2
X_3	Solve stage 2, not stage 1	P_3
X_4	Solve neither stage	P_4

Assuming this multinomial distribution of ability, appropriate ability interaction within a group of four individuals can accomplish a solution even though the group has in it no one member who can solve the problem as a whole; for example, the group whose members symbolically are represented as $X_2 X_3 X_3 X_3$. Consider all possible samples of four ($X_i X_j X_k X_m$) from this population. It is possible to enumerate all groups of four that can interact to accomplish whole solutions solely by pooling their abilities. Any group containing at least individual X_1 , or at least individuals X_2 and X_3 jointly, will be successful. The probability of occurrence of each sample of four is given by the multinomial distribution if P_1, P_2, P_3, P_4 are known. The sum of probabilities of the occurrence of each group of four that can complete a stage-wise solution is the probability of a group solution on the hypothesis of stage-wise pooling of ability. Thus, under Model B, the probability of a group solution is obtained by a special summation of the elements of the multinomial distribution.

Currently, not enough knowledge is available for estimating all the probabilities P_1, P_2, P_3 , and P_4 . At best, in line with current knowledge of the distribution of ability, the psychologist can merely supply reasonable estimates for P_2, P_3 , and P_4 . In Shaw's data, P_1 can be estimated from the sample. This still leaves two degrees of freedom for choices since the sum of the four probabilities is one.

Suppose these two free choices are subject to the restriction that they closely reproduce p_g and that they are not inconsistent with psychological knowledge of the distribution of ability. For the kind of problems treated by Shaw, psychological evidence indicates that the percentage of persons who will fail on both stages will be larger than the percentage who can solve both stages or any one stage. This, of course, does not uniquely determine the four parameters but it is interesting to see that reasonable estimates do exist. For example, if in Shaw's Problem I, $P_1 = .15$ ($p_I = .1428$), $P_2 = .15$, $P_3 = .15$, and $P_4 = .55$, then $p_{gB} = .61$ as contrasted with $p_g = .60$. Here P_2, P_3 , and P_4 were guesses to reproduce the observed p_g . They are also not inconsistent with the distribution of ability. Actually p_g can be reproduced exactly, but it was not considered necessary to alter the p_i 's slightly to accomplish this since enough leeway has already been taken to reproduce a sample value. Moreover, slight changes would not alter any decisions about the reasonableness of the P_i 's. This argument also applies in the following discussion of Problems II and III. Incidentally, $P_2 = P_3 = .15$ leads to

$P_1 + P_3 = P_2 + P_3 = .30$; this indicates that the probability of an individual's success in stage 1 and stage 2 is .30. By (5a), $P_G = .58$ as contrasted with $p_G = .60$, which suggests that the assumption which yields (5a) from (5) is realistic after all.

Moreover, if in Shaw's Problem I, $P_1 = .15$, $P_2 = .30$, $P_3 = .30$, and $P_4 = .25$, a situation definitely inconsistent with the distribution of ability, we get $p_{GB} = .92$, a value noticeably different from $p_G = .60$.

Similarly for Problem III, if $P_1 = .10$ ($p_1 = .0952$), $P_2 = P_3 = .10$, $P_4 = .70$, then $p_{GB} = .42$, as contrasted with $p_G = .40$. Also referring to (5a), $P_G = .35$, as contrasted with $p_G = .40$. In Problem II, $P_1 = .2$, $P_2 = P_3 = .05$, and $P_4 = .70$ yields $p_{GB} = .61$, as contrasted with $p_{GB} = .60$; again referring to (5a), $P_G = .46$, as contrasted with $p_G = .60$. It should be noticed here that this big difference arises from the use of $p_{IA} = .2$, which would lead to a one-stage problem if it were p_I . Notice that this is reflected also in $P_2 = P_3 = .05$, for $P_2 = P_3 = 0$ is a one-stage model. Substitution of the unrealistic observed $p_I = 0$ would yield nonsensical results. This information is presented in Table 2.

It is interesting to note the premium gained by the two-stage model. Model B can be made to account for most of the excess ($p_G - p_{GA}$) not accounted for by Model A. If Model B holds, the excess is the probability of a group solution when individual X_1 is not in the group of 4. For the weights described, this is .13 for Problem I, and .077 for Problem III; these should be compared with $(p_G - p_{GA}) = .14$ for Problem I, and .07 for Problem

TABLE 2

	Problems		
	I	II	III
p_G	.60	.60	.40
p_{GA}	.46	.00	.33
p_{GB}	.61	.61	.42
P_1	.15	.20	.10
P_2	.15	.05	.10
P_3	.15	.05	.10
P_4	.55	.70	.70

p_I = ratio of individual solutions to attempts

p_G = ratio of group solutions to attempts

p_{GA} = estimate of P_G from Model A and observation p_I

p_{GB} = estimate of P_G from Model B and weights P_1, P_2, P_3 , and P_4

P_1 = probability that an individual will solve both stages in Model B

P_2 = probability that an individual will solve stage 1 but not stage 2 in Model B

P_3 = probability that an individual will solve stage 2 but not stage 1 in Model B

P_4 = probability that an individual will not solve either stage in Model B

III. For Problem II, the weights used lead to an excess of .017, but this is just another reflection of the fact that the replacement of p_I by p_{I_A} leads to a one-stage problem.

The stage-wise model hypothesizing the pooling of ability tends to reproduce the observed p_G when reasonable weights are used. Indeed, unreasonable weights produce major discrepancies from the observed p_G . The implication of the model is that group superiority may be conceived as a function only of pooling the abilities of its members. Ultimately, empirical estimates must be obtained for P_2 , P_3 , and P_4 . One experimental procedure for such estimates would require individuals to solve the problem. For instance, in a two-stage problem, those individuals solving the problem (as in the Shaw data) provide a basis for estimating P_1 . Some individuals who failed the whole problem, however, will have accomplished stage 1 successively but failed on stage 2, providing a basis for estimating P_2 . The remainder, those who could not accomplish stage 1, would be given the problem reduced by the accomplishment of stage 1, reported as a fact, with the requirement that the "new" problem be solved. Some of the individuals will then solve the "new" problem providing a basis for estimating P_3 .

When P_1 , P_2 , P_3 , and P_4 are estimated by p_1 , p_2 , p_3 , and p_4 on the basis of sample observations, assuming Model B holds, a value p_{GB} will be obtained and contrasted with p_G . As in Model A, the probability that an observed difference will be exceeded by chance must be computed in order to examine the tenability of the model. Under the assumption that Model B holds, and replacing P_1 , P_2 , P_3 , and P_4 by their estimates, it is possible to obtain the exact distribution of $(p_G - p_{GB})$, although it is extremely tedious to compute. If p_i is based on n_i observations, then p_{GB} can assume $(n_1 + 1) \cdot (n_2 + 1) \cdot (n_3 + 1) \cdot (n_4 + 1)$ values. Even if the sample sizes are small, say $n_i = 5$, p_{GB} takes on 1296 values. This, plus the difficulty of actually computing the probability of a difference $(p_{GB} - p_G)$, renders the technique somewhat useless. Moreover, for large samples an asymptotic method seems fruitless because of the special way the multinomial distribution is summed for this situation.

Suppose, however, a confidence interval for P_G , say P_{GL} and P_{GV} is obtained from p_G . Assuming the model holds, all the sets p_1 , p_2 , p_3 , p_4 which yield values between P_{GL} and P_{GV} inclusive, form a confidence region for P_1 , P_2 , P_3 , P_4 . Actually all that need be done then is to consider the value p_{GB} yielded by the observed p_1 , p_2 , p_3 , p_4 . If this value lies between P_{GL} and P_{GV} the model is tenable for the specified confidence coefficient employed, let us say $1 - \alpha$, or equivalently for the significance level α .

Pooling of Data

Shaw pools the results for the three problems, neglecting the fact that the same individuals and same groups worked the three problems in the same

sequence. Thus, she contrasts 8/15 or 53 per cent success for groups with 5/63 or 7.9 per cent success for individuals. Using the z test given by (1) with the awareness that the lack of independence renders it inadequate, this difference is statistically significant at the 5 per cent level. Moreover, since the correlation between observations can be assumed to be positive, the decision of statistical significance is on the conservative side. Also, Model A is rejected using the z test given by (4). It should be emphasized that of the five groups, two solve none of the three problems and two solve all. Of the twenty-one individuals, none solves more than one of the three problems! Two alternate hypotheses are suggested: 1) Model B is operating; 2) groups do better than individuals in a sequential solution of problems of the same kind. Hypothesis 2 can arise from three possibilities: (a) negative transfer in individuals, zero or positive transfer in groups; (b) zero transfer in individuals, positive transfer in groups; (c) positive transfer in individuals, greater positive transfer in groups. As regards hypothesis 2, Cook (1), using two versions of the disc problem (Problem III), varying in difficulty of sequence, implies "that transfer 'spuriously' lowers the probability of a given individual achieving the same degree of success or failure (relative to the rest of the groups) on both problems." The evidence from Shaw's groups suggests somewhat the same conclusion by indicating the plausibility of positive transfer in groups in sequential solution of problems of the same kind. A carefully designed experiment to ascertain the superiority of groups over individuals in transfer of training is suggested by this combined evidence.

REFERENCES

1. Cook, T. W. Amount of material and difficulty of problem solving. II. The disc transfer problem. *J. exp. Psychol.*, 1937, 20, 288-296.
2. Eisenhart, C. Inverse sine transformation of proportions. In Eisenhart, Hastay, and Wallis, *Techniques of statistical analysis*, New York: McGraw-Hill, 1947.
3. Kasner, E. and Newman, J. *Mathematics and the imagination*. New York: Simon and Schuster, 1940.
4. Lorge, I., Fox, D., Davitz, J., Weltz, P., Herrold, K. Products of individual and of group: Review of literature. Unpublished manuscript developed in connection with Contract AF 33 (038) 28792, Human Resources Research Institute, Maxwell Air Force Base, Alabama.
5. National Bureau of Standards. *Tables of the binomial probability distribution*, Applied Mathematics Series 6, January 1950.
6. Paulson, E. and Wallis, A. Planning and analyzing experiments for comparing two percentages. In Eisenhart, Hastay, and Wallis, *Techniques of statistical analysis*, McGraw-Hill, 1947.
7. Shaw, M. E. Comparison of individuals and small groups in the rational solution of complex problems. *Amer. J. Psychol.*, 1932, 44, 491-504.

Manuscript received 5/17/54

Revised manuscript received 9/16/54

ON THE DESIGN OF AUTOMATA AND THE INTERPRETATION OF CEREBRAL BEHAVIOR

STANLEY FRANKEL

CALIFORNIA INSTITUTE OF TECHNOLOGY*

In principle it is possible to design automata to display any explicitly described behavior. The McCulloch-Pitts "neuron" is a convenient elementary component for the control mechanisms of automata. Previously described techniques permit the design of an automaton which would arbitrarily well simulate human behavior. The difficulty of producing such a design lies primarily in formulating an explicit description of the required behavior. The control mechanism of such an automaton would be of very great logical complexity. Its mode of operation probably would not resemble that of a human brain. The brain is more plausibly represented by stochastic models as proposed by Hebb. Such models can more easily be designed or understood by reason of lesser logical complexity. A method of computational investigation of the functioning of such stochastic models is described. Several extremely simple models have been investigated. One is shown to have properties suggestive of learning ability.

I. *Introduction*

The possibility of constructing or designing devices which can to some extent simulate the complex patterns of behavior of man or the higher animals, particularly those aspects of behavior favored by an intact nervous system, has long excited lively speculation. It seems desirable to continue such speculation in the present era and to bring modern resources to bear on the problem in the hope that some light may be shed on the mechanisms operative in these complex behaviors.

At best it can be hoped that this approach may yield hints of possible mechanisms; yet, in light of the formidable difficulties of direct studies of complex nervous systems, even this modest hope amply motivates such investigations. The utility of this approach need not be further stressed since many authors have testified to its fruitfulness.

A number of lines of argument directed toward the establishment of limits on the range of possible behaviors possible for an automaton have been explored. Several often-proposed arguments to this end have been shown by Turing (14) to be incapable of clear formulation or incapable of leading to the desired limitation. These arguments seem to be stimulated by a common motive: in the course of a normal childhood development a person finds it possible to some extent to reduce to order the initial chaos of observation of the external world, including animate objects. By reason of

*Present address: 30792 Driftwood Dr., So. Laguna, Calif. The author is indebted to many friends for helpful discussion and criticism, particularly to Miss Winifred Whitfield and Dr. John von Neumann.

this orderliness he finds himself possessed of a measure of control. The subjectively recognized "self" is somewhat separated from the environment in the ordering. Since the analogy between his objectively observed self and his companions is too striking to be overlooked, there arises the desire to exempt his kind from the observed lawfulness of the external world. (The term "kind" by intent lacks precision. Primitive peoples have often extended this exemption very generously. The modern tendency seems to be to restrict it to our own species or, more narrowly, to one's own tribe, sex, sect, etc.) A view which may be so motivated is aptly expressed by Jefferson (6): "No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, feel grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or depressed when it cannot get what it wants."

That an upper bound can in fact be placed on the complexity of behavior which an automaton could display can be shown by paraphrasing an argument due to Turing (13). He describes as a *computable number* a denumerable sequence of digits which can be generated by a computer of finite complexity (i.e., an infinite ordered sequence of decisions implicit in a description which is finite, however lengthy). He shows that the set of computable numbers is denumerable; thus they are negligibly few among all real numbers. In a similar vein we may approximate the entire range of environmental circumstances and the reaction of an automaton thereto by denumerable sets. A *realizable behavior pattern* is a functional relation between these sets which permits an implicit description in finite terms, hence demonstrable by an automaton of finite complexity. Turing's argument, appropriately transposed, shows that the realizable behavior patterns are denumerable, hence include negligibly few among the continuum of behavior patterns. In outline, this argument may be stated as follows: The finite descriptions may be translated into a standard language in which each description consists of a finite sequence of words drawn from a finite vocabulary. These descriptions may be classified by number of words and, within each class, arranged in alphabetical order. Thus the set of all descriptions which the language admits (including meaningless and redundant descriptions) can be enumerated. It follows that the set of behavior patterns which can be so described is also denumerable.

Unfortunately (from the point of view of the desire described above) this argument will not serve to limit the extent to which an automaton can mimic human behavior unless it can be shown that the range of human behavior patterns extends beyond this denumerable set.

It must have appeared likely in previous centuries that a quite stringent limitation on the possibility of actually constructing an automaton to mimic human behavior is imposed by grossness of constructable elements of machinery—gears, levers, pulleys, etc.—in comparison with human size. This limitation need not be seriously considered here, since it is of little importance

to the present purpose that an automaton *look* like a man. In any case, this appearance of limitation is weakened by the present development of semiconductor electrical elements, which promises almost unlimited miniaturization of the types of electronic apparatus which now seem suitable elements for the construction of automata.

It is also to be noted that valuable hints to neurophysiology may arise from the design of an automaton which, by reason of technical or economic limitations, may not be constructed in the metal. However, as the example of Walter's Testudo (16) strikingly displays, the verisimilitude of an automaton's simulation of animal behavior can far better be judged by direct observation of the behavior of the automaton than by study of its wiring diagram or differential equations. It thus seems desirable to give preferential, although not exclusive, consideration to those designs for automata which could be built without extravagant effort.

Another line of argument bearing on the range of behavior patterns accessible to man-made automata has been explored by von Neumann (15). It is clear that the complexity of the behavior pattern of an automaton is subject to an upper limit dependent on the complexity of its mechanism. This complexity in turn would appear to be limited by the extent of human ingenuity, which may in turn be similarly limited. One is at first tempted to believe that suitable measures of these complexities would permit proof of a hierarchic ordering of machines. Further, he might be tempted to believe that a machine can fabricate, or in some sense design or conceive, only machines of lesser complexity than itself; similarly he might think that a man-made automaton must be, in respect to this measure, inferior to its creator. It may seem that some degradation of information must occur between the construction and operation of a machine. The hope for such a theorem is damped by von Neumann's description, in outline, of a machine capable of fabricating a duplicate of itself after first (this to exclude trivial solutions) building a locomotive. The trick of design permitting this description depends on the distinction between the *actual* and *logical complexities* of a machine. For example, the elaborate set of instructions which governs its manipulations may be carried as a perforated tape which, though of great actual complexity, can be copied by reiteration of a simple elementary operation.

One might still hope to show it to be impossible for a man to *understand* a device of a complexity equal to his own even though he could build it (suitable meanings being given to these terms). Such a demonstration would require comparisons of *logical complexity*, so defined as not to increase markedly with simple reduplication of components, rather than *actual complexity*, defined, for example, by a simple counting of parts. The description of a *universal computer* given by Turing (13) would seem to weigh against this hope. This instrument, of finite *logical complexity*, is one capable of

predicting the operation of any computer, of however great complexity. This great reduction in complexity is obtained at the expense of speed of operation which, though desirable, can hardly be regarded as of fundamental significance in human understanding. Moreover, an increase in speed can often be obtained by duplication of components without increase in logical complexity. It thus does not seem likely that arguments along these lines can place the desired limit on the range of behavior possible for an automaton.

II. *Neural Network Models*

Automata have variously been conceived as primarily composed of marble, clay, clockwork, and, more recently, of vacuum tubes, relays, etc., (2). The chief advantage of the modern components lies in the ease with which they can be physically assembled in practically operative equipment. If only a theoretical reduction to practice is required, the advantage lies rather with the more intuitively understandable cams, detent gears, levers, and Jacquard cards involved, e.g., in Babbage's conception of his *analytical engine* (1). A still more convenient basic component is the idealized "neuron" (designated as the M. P. neuron) used by McCulloch and Pitts (8), Culbertson (3), and others. The properties of this mathematical construct are sufficiently simple to be easily realized in manufacturable equipment, yet bear a close analogy to the observed neurons of physiology.

Briefly, the properties of a network of these mathematical neurons are the following: In each of a series of time intervals each neuron may "fire" or remain quiescent. Two neurons may be connected by a process which produces an effect on the second neuron whenever the first fired in the preceding time interval. This effect may tend either to excite or to inhibit the present firing of the second, depending on the kind of connecting process. The firing or non-firing of each neuron is dependent on the number of presently received effects of the two kinds. The functional dependence of this decision on the two numbers is subject to choice.

The mathematical neuron is well adapted to the design of automata. One can, with surprising ease, design circuits to mediate even quite complex activities (3). The chief difficulty lies in formulating a precise description of the desired activity. The process of translating this description into circuitry can be carried out essentially by rote. It is thus not out of reason to assume that circuits could be designed to mediate each identifiable aspect of the behavior of a human adult. It may further be assumed possible to design suitable interlocks which suppress the activities of these circuits until after appropriate environmental circumstances have occurred. For the purpose of the argument to follow it will be assumed that in this way one could design an automaton which would satisfactorily simulate both the learning steps and the learned behavior patterns of an entire lifetime. By reason of its part-by-part design this may be termed a *block-diagram* model.

III. *Defects of the Block-Diagram Model*

The neural-network-controlled automaton so designed seems at first to supply a complete answer to the primary question of this investigation. It would simulate human behavior in every situation taken into consideration by the designer; thus, the completeness of the simulation is limited only by his patience. Yet for a number of reasons this answer is unsatisfying; this block-diagram automaton does not seem to present a close analogy to the observed human nervous system as indicated by the following considerations:

(1) The automaton might include some circuits permitting it, say, to display a full command of every modern language and further circuits inhibiting their action until environmental circumstances (e.g., exposure to a course of study of a language) make the display of each ability seemly. This heroic design effort might require the use of most of the 10^{10} -odd neurons permitted by direct analogy, yet would still not provide, for example, for the learning of Sanskrit. The view that human ability to learn any few of the enormously many known languages is based on the release from inhibition of precisely arranged circuits is strikingly unappealing.

(2) The learning ability of the higher animals seems quite unsystematic in comparison with that of an automaton designed in this way. For example, a man's capacity to learn to drive an automobile or a rat's to learn a T-maze can hardly be ascribed to the pressure of natural selection in the short period since the introduction of these features of environment. They rather suggest the operation of an unspecific learning ability operative in a wide range of circumstances. An automaton which mimics the behavior of a laboratory rat by means of many marvelously contrived circuits, each initially frustrated by an equally marvelous inhibiting circuit, suggests great virtuosity but not true efficiency of design.

(3) If the block-diagram automaton is efficiently designed with respect to the number of neurons used, the modifications in its behavior which would result from the extirpation of a small fraction of its neurons would be most striking. Some of its possibilities of learning would disappear, other abilities would spring forth full-blown as their inhibiting mechanisms are inactivated, without the normally required training program. Some abilities previously acquired by training would be irrevocably lost. If the control circuits are redundantly designed, reliability of operation being achieved at the cost of a many-fold increase in the number of neurons used, similar effects would follow the destruction of larger parts of the control circuits.

The effects of cerebral injury in the mammals present a quite different picture (10). The resulting changes in behavior are often surprisingly mild, suggesting considerable redundancy of design; to some extent these changes are reparable by retraining. These changes are, from the present point of view, uniformly pejorative. [As Wiener points out (17), prefrontal lobotomy

may usually be expected to increase a patient's tractability, not his wit.] To find, for example, the victim of a brain accident subsequently in command of a new language would occasion surprise.

(4) The complexity of the ingeniously contrived neural circuit controlling the behavior of the block-diagram automaton might be comparable with that of the intricate interconnection of the 10^{10} -odd neurons of a human central nervous system. The latter, however, is presumably built to a pattern held in the 10^5 -odd genes controlling human heredity. [The number of human gene positions has variously been estimated as 25,000 to 100,000 (12).] These must also be presumed to determine the architecture of other than neural tissues and much of intracellular physiology as well. Thus, it seems likely that the genes serving to determine the circuitry of the human central nervous system number no more than a few thousand. This consideration suggests that the essential *logical complexity* of the human nervous system is far less than the maximum which the number of neurons and synaptic junctions would permit.

IV. *The Hebb Model*

A very different model of a complex central nervous system has been examined by Hebb (5). (Hebb does not use our present oblique approach, but addresses himself directly to the study of actual nervous systems. For this discussion, it is convenient, however, to regard his description as that of a proposed automaton design.) The chief design aim of this model is minimization of logical complexity. The elements composing his model are given many of the properties which the cells of the human nervous system are currently thought to have. They may thus be termed neurons less metaphorically than the elements described above. They have one further property which, though it finds some support in neurophysiology, may be regarded as an *ad hoc* assumption. It is assumed that each firing of a neuron is accompanied by the strengthening of the synaptic junctions through which it was stimulated. This property, called *neurobiotaxis*, makes any often-repeated chain of neural firings progressively more easily initiated.

The Hebb neuron is a much more complex structure than the M. P. neuron. It partakes in a few hundred rather than in a few synapses. Its firings are determined by a complicated interplay of stimulations, periods of continuously varying excitability following a previous firing (the relative refractory period); its firings are not subjected to a coarse-grained time quantification. Moreover, it has, by reason of the assumed neurobiotaxis, a more fine-grained and more retentive memory than the one time unit memory of the M. P. neuron. This greater complexity of the Hebb neuron does not give added scope to the organizations which can in principle be based on it, since its properties can be duplicated with any desired accuracy by structures composed of M. P. neurons. The Hebb automaton thus relegates greater

logical complexity to its basic element than does the block-diagram model.

The logical simplicity of the Hebb model lies in the bold assumption that the interconnections of the neurons are for the most part *not planned*. The neurons are regarded as produced in vast numbers by a broadcast mechanism (e.g., by successive cell division) and to position and interconnect themselves in a way which is determined by design only as regards gross architectural features. The detailed wiring of the model, in which neurons form synaptic junctions with others, is randomly determined. Here and in what follows the term "random" is used in the lay sense—unplanned, nondescript, determined by happenstance—rather than in the broader mathematical sense. The Hebb picture of the cerebral cortex may be likened to the subsoil aspect of a forest. The complex matting of roots is not the result of meticulous engineering but of the chance placement of the grains of sand, drops of water, etc., which influenced the growth of each of the roots. There are architectural features—tree roots by and large go deeper than grass roots—but the precise configuration of roots is not subject to plan. This is not meant to suggest that the growth of a grass root, or of an axon, is exempt from causality, but only that myriad other configurations of roots would serve as well to nourish and support the forest.

The essence of Hebb's discussion lies in the observation that a large random network of neurons must be presumed to include many circular (reverberatory) chains, capable of sustained activity when once excited. Those reverberations frequently excited by particular combinations or concatenations of stimuli tend to be fixed by neurobiotaxis and may be evoked by progressively smaller aliquots of the constellation of stimuli initially required. The first-formed elementary reverberations will interact among themselves to form higher-order associations and combinations, thus leading to a complex hierarchical structure.

In a network of M. P. neurons provided only with excitatory interconnections, a stimulus can more readily excite an appropriate response than suppress other inappropriate responses. [Von Neumann (15) has shown that networks of M. P. neurons can be given full logical universality without the use of inhibitory interconnections. The device used, however, does not lend itself to use in a random model.] The Hebb neuron, unlike the M. P., displays a significantly long refractory period. This makes possible the suppression or inhibition of reverberatory activity by excitatory processes alone. If two reverberatory chains share the use of a number of neurons, the excitation of one reverberation may, by fatiguing shared neurons, tend to suppress activity in the other. Similarly, any strikingly intense, widespread excitation of the network, which may be identified with painful stimuli to an animal, will tend to break up the current large-scale pattern of activity. Hebb's model looks to this disruption of over-all patterns of activity by intense stimulation to effect macroscopic (goal-directed) learning.

V. Possibilities of Computational Investigation

Hebb's extensive qualitative discussion has the aim of making plausible the view that the impressive abilities of a human nervous system are explicable with this (conceptually) simple set of hypotheses. The above discussion is intended as a description of this aim, not as a summary or as a critical review of Hebb's argument. The success of such a plausibility-proof must be judged by each reader for himself. I find it profoundly convincing.

The difficulties which stand in the way of a firm analytic proof of the adequacy of the Hebb model are highly formidable. These do not, as might first be thought, have primarily to do with the enormously large number of chaotically interconnected components. Statistical techniques for investigating the properties of such assemblages are reasonably well developed. The chief difficulties of analysis lie rather in the moderately great complexity of the elementary unit and of the gross structural features of the observed human nervous system. In order to preserve reasonable verisimilitude, a model of the human nervous system might require the description of, say, one hundred distinct regions, each with its own statistically described pattern of organization. This number is not so small as to permit carrying out, with reasonable labor, an analysis which takes each region separately into account, nor is it so large—and each region so unimportant—as to permit using a statistical description which overrides their distinctions.

Despite the formidable difficulties confronting an attempt to prove the adequacy of the Hebb model as representing the human nervous system, it does not seem out of reason to attempt more modest checks of the basic features of the model. Considerable simplification of the task can be effected by omission from the model of many features which are auxilliary to the dramatic and characteristically mammalian behavior patterns. It would seem reasonable to omit from a preliminary analysis any representation of the neurological components of the homeostatic mechanisms controlling temperature, pH, etc. Similarly, the model might be divorced from most, if not all, of the normal mammalian sensorium and control over musculature. Some means must be provided for representing an interaction of the model with its environment, but for first analysis it would seem sufficient to provide some logically simple (though unphysiological) input and output mechanisms. There would likewise seem little reason to provide the model with any wired-in interconnections between input and output to provide for the demonstration of unlearned reflexive or instinctual behavior. In brief, the model need not be required to display any of the aspects of animal behavior for which the possibility of mechanical representation is subject to little doubt.

So stripping the model of lesser requirements may considerably simplify checking its ability to simulate intelligence. If carried to completion, however, this stripping might prevent the display of any goal-directed learning.

A rat who feels no hunger cannot be expected to display shrewdness in a food-goal maze. The simplification of the model should thus stop short of the removal of all affective inputs. It may suffice, however, to leave one goal-associated feature of the input which, in the interpretation of the behavior of the model, plays the role of a generalized indication of pain (or alternatively of pleasure). Each particular environment of the model will be specified by a functional dependence of the inputs to the model on its (present and prior) outputs, i.e., its experiences are at least in part determined by its behavior. The extent to which the behavior of the model serves to diminish (or, in the alternative case, to increase) the frequency of stimulation of the goal-associated input is then a measure of the goal-directed learning displayed by the model in each environmental situation.

Even with these simplifications it is not clear that the present techniques of mathematical analysis permit a more penetrating study of the expectable performance of stochastic neural network models than that to be found in Hebb's qualitative discussion. A more promising approach would seem to be offered by the use of a modern general-purpose digital computer to simulate the behavior of specific examples of the model in specific environmental situations. Some loss of mathematical rigor is unavoidable in this method of study since the performance of a few haphazardly selected examples would be taken as typifying the performances of the enormously large ensemble of possible realizations of each model, the details of which are randomly specified. The danger of being misled by a fluke performance is, however, no greater than that which occurs in most stochastic investigations and admits the usual statistical safeguards. This technique of investigation of a highly multidimensional ensemble by sampling, known as the Monte Carlo Method, has been investigated by Ulam, von Neumann, and others (9). In some applications this technique proves notably efficient (18). The study of the Hebb model and other stochastic neural network models appears to be such an application.

VI. *The Three-Layer Model*

Prior to his acquaintance with the Hebb model, the author initiated the investigation of a stochastic neural-network model composed of M. P. neurons (17). In this three-layer model the neurons are distributed upon a surface in three classes (layers) serving particular purposes. One, the trunk layer, was to transmit to all parts of the surface notice of the reception anywhere of a painful stimulus. Another, the granular layer, was to record certain special events by the initiation of spatially localized reverberations. These reverberations may be extinguished by the passage of a wave of excitation along the trunk layer. They thus provide a pain-limited temporary memory of the occurrence of the special events which initiated them. In the third, the primary layer, the neurons are interconnected by long-range pro-

cesses, unlike those of the trunk layer and granular layer, which have only local connections. The simultaneous firing of two neighboring neurons of the primary layer constitutes the special event to be recorded in the granular layer. A reverberation in the granular layer is to produce a progressive increase in the sensitivity of neurons of the primary layer near to the neurons which initiated the reverberation.

It was hoped that by appropriately specifying the statistical structure of its neural interconnections the network could be shown to display properties suggestive of learning ability. This learning ability of the model may be expected to increase indefinitely (at least in some quantitative sense) as the size, but not the logical complexity, of the network is increased, hence without increase of the ingenuity invested in its design.

It proved easy to show the desired properties for the trunk and granular layers (4). The trunk layer neurons were densely interconnected and given an appreciable refractory period. These parameters could be widely varied, still permitting the propagation of a wave of excitation. The neurons were arranged in a rectangular array, opposite edges of which were regarded as contiguous so as to make the surface topologically of spherical or, more conveniently, of toroidal character. In this way disturbances owing to atypical characteristics at boundaries were avoided. The wave of excitation would spread over the entire surface leaving behind a refractory zone and, upon reconverging to a point, be extinguished.

A few possible structures for the granular layer were examined. The performance of the layer was found to be favored by a short refractory period and by considerable statistical fluctuation in the degree of local connectivity; hence the name granular. This permits the maintenance of many independent local reverberations while preventing any long-range spreading of excitation. An extreme, perhaps trivial, solution is to give each neuron unit threshold and one self-exciting process.

The study of the properties of the primary layer presented considerably greater difficulty and seemed to require the use of calculating equipment of greater speed and flexibility than was readily available. It was proposed at that time (1949) that an electronic calculator of considerable memory capacity be used in the further exploration of a three-layer model, with attention focused chiefly on the primary layer.

In the summer of 1951, through the generosity of the National Research Development Corporation and of the Department of Mathematics of the University of Manchester, the author had opportunity to initiate this exploration with the use of the Manchester Mark I computer. The results of the calculations made at that time are described below. Although they are of a qualitative and preliminary nature they support the suggestion that this technique of investigation is likely to prove fruitful.

VII. *Manchester Calculations*

It was decided, for the sake of computational simplicity, to make no attempt to represent explicitly the neurons of the granular and trunk layers but to subsume their supposed effects on the thresholds of excitation of the primary layer neurons in the rules governing the computations.

In a first series of calculations planned, a large number of neurons was to be represented. Each was to receive excitatory processes from a fixed or variable number of others selected by a random process. To avoid overburdening the rapid-access memory capacity of the computer, it proved convenient to represent the series of random numbers which describe the interconnections of the neurons by an algebraic formula from which they could be repeatedly calculated rather than to store the series in the memory. These numbers are thus not truly random, but have sufficient complexity to be considered quasi-random, i.e., sufficiently disorderly for the purpose.

In each cycle of the calculation the computer determines and displays which of the neurons fire; information for use in the succeeding cycle is recorded. The firing of a neuron is determined by the number of neurons, among those from which it receives excitatory processes, which fired in the preceding cycle. If that number equals or exceeds its assigned threshold it fires, otherwise not. Again to avoid overburdening the computer memory it was at first planned to take no account of the number of cycles elapsed since the neuron last previously fired, i.e., not to assume a refractory period exceeding one cycle. Modifications in the behavior of the network were to be effected by changes in the thresholds.

As a preliminary to these experiments a series of calculations was carried out to determine suitable ranges of the number or mean number of excitatory processes brought to each neuron and the constant or mean threshold. It soon became evident that no suitable values of these parameters could be found. For any value of the threshold exceeding unity the level of reaction was intrinsically unstable. The number of neurons firing either fell to zero after a few cycles or rose to almost the full number of neurons represented. An elementary statistical calculation shows that this behavior is to be expected in the models as tried.

Two methods of overcoming this difficulty presented themselves. One is the use of neurons with definite refractory periods considerably exceeding the synaptic delay time, i.e., many cycles of the calculation. This should tend to depress the upper stable level of reaction by reason of the refractory condition of most of the neurons at each cycle. This procedure seems unattractive, since the firing of a neuron would then depend chiefly on its release from inhibition rather than upon the immediately preceding pattern of firings.

The second method of stabilizing the level of reaction is logically appealing though seemingly unphysiological. It is to use only inhibitory rather than excitatory interconnections of the neurons. This gives the level of reaction negative rather than positive feedback characteristics, thus producing a single stable level of reaction. This procedure avoids the necessity of maintaining in the computer memory a record of the number of cycles elapsed since the last previous firing of each neuron, as would be required if long refractory periods were taken into account. It was accordingly decided to adopt this latter procedure.

VIII. *The Linear Inhibitory Model*

Having taken this one step away from physiological plausibility, another became appealing. The complexity of the computational procedure can be considerably reduced by limiting the group of neurons to which each neuron may be responsive. Accordingly, it was decided to represent the neurons of the primary layer as arranged in a circular sequence and to select the neurons sending inhibitory processes to each neuron from among the forty immediate predecessor neurons in this sequence. A further computational simplification is afforded by making the firing of each neuron contingent on the just prior computed firings of its predecessors rather than those of the preceding cycle. It is to be noted that this simplification is effected at the cost of a great reduction in the amount of information in storage at each stage.

The computational procedure, so simplified, was set up with the following characteristics: Each neuron received inhibitory processes from some among its forty predecessors, each of these being included or excluded with equal probability. Initially each neuron was given a threshold of inhibition of five, i.e., if five or more among its twenty-odd selected predecessors had fired it did not fire; otherwise it did. Arrangements were provided to replace this determination by selected firings of some sensory neurons to represent environmental stimuli. This arrangement may be expected to lead to the firing of approximately one-fourth of the neurons in each cycle. A cycle now becomes simply one traversal of the circular array of neurons from an arbitrary starting point. Provision was also made for the increase by unity of the thresholds of all neurons which had fired in the preceding cycle, at the choice of the operator. He was thus enabled to simulate the assumed effect of the granular and trunk layers in rewarding a successful performance.

This linear inhibitory model performed qualitatively as expected on its initial cycles. The number of neurons firing per cycle remained approximately constant. No pattern was readily visible in the change from cycle to cycle in the set of neurons fired. Thus, prior to any learning experiment the activity of the model seemed substantially random. It is to be noted, however, that the state of the model at any moment is specified by only forty binary

alternatives—the firing or not of the forty preceding neurons. Taking into account the fact that in the mean only one-fourth of the neurons fire this represents a store of only $33\frac{1}{2}$ bits of information. It is therefore to be expected that the appearance of randomness of the firing pattern is not deep-seated.

As a first simple learning experiment it was decided to promote the firing of a particular neuron, making no use of the input mechanism representing environmental stimuli. A neuron was chosen which, prior to the learning experiment, fired in very nearly one-fourth of the cycles. Its firing was taken as the criterion of a successful cycle. In the learning experiment the model was run as initially set up until the first cycle in which the selected neuron fired. The model was then rewarded as described above, i.e., each neuron which had fired in that cycle had its threshold raised to six. On superficial examination the experiment seemed strikingly successful; there-after the selected neuron fired in every cycle, although no further reward was supplied! On closer examination, however, it appeared that the structuring of the firing pattern resulting from the reward was excessive. After the reward the model fell into an immutable pattern, each neuron either fired in every cycle or in none. This fixed pattern was similar to that which occurred in the rewarded cycle but not identical. Thus, if the selected neuron had been one for which the two patterns differed, the experiment would have seemed totally unsuccessful; the model would have learned the opposite from the intended behavior.

The result of this experiment is unsatisfactory in another way. Since the firing pattern was fixed by the first reward, the model was not capable of further learning. A milder form of reward would presumably have been preferable in diminishing the likelihood of fixing the wrong pattern of behavior. It is also clear that a satisfactory model requires a much greater stock of randomness in its initial behavior.

This result suggests that a learning mechanism may need to be guarded against excessively rapid learning, which could lead to its leaping to unjustified conclusions. The optimum learning rate would seem to be determined by the opposing hazards of accidental learning, brought about by statistical fluctuations and accidental correlations of actually unrelated things, and the dangers (depending on the environmental circumstances) of learning too slow. It is interesting to speculate on the effect of the considerable increase in the prevalent life span which the human species has experienced in the course of its last few thousand generations. It seems possible that this has made more prevalent the defect of human intelligence that arises from too-rapid learning. It may appear that the typical mentally maladapted individual has not learned too little but rather too much that is not true.

It has not as yet proved possible to test the behavior of this model as

modified in the ways suggested by this result. It is the author's hope that the suggestion of success shown by this model will serve to stimulate similar investigations in some laboratories having electronic computing machines.

REFERENCES

1. Babbage, H. P. Babbage's calculating engines. London: E. and F. N. Spon., 1889.
2. Chapuis, A. Les automates. Neuchatel: Edmond Droz, Editions du Griffon, 1949.
3. Culbertson, J. T. Consciousness and behavior. Dubuque: W. C. Brown Co., 1950.
4. Frankel, S. Proposed neural network investigation. Dig. Comp. Gr., Calif. Inst. Tech., 1949.
5. Hebb, D. O. The organization of behavior. New York: Wiley, 1949.
6. Jefferson, G. The mind of mechanical man. *Brit. med. J.*, 1949, 1, 1105.
7. Jeffress, L. A. (Ed.) Cerebral mechanisms in behavior. New York: Wiley, 1951.
8. McCulloch, W. W. and Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bul. math. Biophysics*, 1943, 5, 115-133.
9. Monte Carlo method. Nat. Bur. appl. math. Ser., No. 12, 1951.
10. Nielsen, J. M. Agnosia, apraxia, aphasia. New York: Paul B. Hoeber, 1946.
11. Northrop, F. S. C. The neurological and behavioristic psychological basis of the ordering of society by means of ideas. *Science*, 1948, 107, 411-417.
12. Spuhler, J. N. On the number of genes in man. *Science*, 1948, 108, 279-280.
13. Turing, A. M. On computable numbers, with application to the *entscheidungsproblem*. *London math. soc. Proc.*, s. 2, 1936, 42, 230-265.
14. Turing, A. M. Computing machines and intelligence. *Mind* (Oct. 1950).
15. Von Neumann, J. Probabilistic logics. (Notes R. S. Pierce) Dept. of Math., Calif. Inst. Tech., 1952.
16. Walter, W. G. An imitation of life. *Sci. Am.* 1950, 182, 42-45.
17. Wiener, N. Cybernetics. Cambridge, Mass.: Technology Press, 1948, p. 173.
18. Wilson, R. R. Monte Carlo calculation of showers in lead. (Abs.) *Phys. Rev.*, 1950, 79, 204.

Manuscript received 4/12/54

Revised manuscript received 7/20/54

BOOK REVIEWS

J. P. GUILFORD. *Psychometric Methods*. (2nd Ed.) New York: McGraw-Hill, 1954, pp. ix + 597.

Psychometric Methods has been renovated and enlarged for its second edition. Although it retains the character of its 18-year old predecessor, the new edition includes some major changes. An introductory chapter on measurement theory has been added. Most of the statistical topics have been removed, including the old chapters on simple and multiple correlation. The treatment of psychophysics and scaling has been improved by forming chapters on psychophysical theory and on principles of judgment from material formerly interspersed in the descriptions of the specific methods. Psychological testing now occupies three chapters instead of one. Finally, the problems at the end of each chapter have been revised and are accompanied by answers whenever practicable.

The book begins with a lucid account of the logical basis of psychological measurement, and a discussion of nominal, ordinal, interval, and ratio scales. This is followed by a comparison of the classical psychophysics of Weber ratios, difference limens, and Fechner's law, with the modern psychophysics of discriminial dispersions, the law of comparative judgment, and stimulus-response matrices. (*S* is used for stimulus and *R* for response, instead of the awkward reversal perpetrated by the Germans.) The third chapter covers mathematical functions, curve fitting, and probability distributions. The major psychophysical methods and scaling methods are covered in the next seven chapters. Included are the methods of average error, minimal changes, constants, pair comparisons (Guilford prefers *pair* to *paired*), rank order, equal sense distances (bisection), equal-appearing intervals, fractionation, constant sums, and successive categories. Experimental designs and computational procedures are illustrated for each method, the pros and cons are discussed, and variants of the methods are noted in short paragraphs. Short sections are also devoted to allied problems, including multidimensional scaling, the objectivity of judgments, and the prediction of first choices. The scaling material concludes with chapters on rating scales and principles of judgment. The latter includes discussions of judgment times, the time-order error, anchoring, judgment sets, regression phenomena, and Helson's concept of adaptation level.

The field of testing requires three long chapters. The discussion of test theory includes a detailed account of the theory based on independent true and error scores, and brief mention of some new additions or alternatives proposed by Lord, Loevinger, Ferguson, and others. Speed and power problems and scoring problems are discussed. Reliability, validity, and item analysis are treated at length. A brief account is given of attitude scale construction. The final chapter is devoted to factor analysis. It includes discussion of general issues in factor analysis, and provides detailed recipes for centroid factoring and graphic rotation of axes.

It is not possible to encompass all of present-day psychometrics in a single volume. However, Guilford has managed to include most of the popular techniques, and has provided extensive references for those who want supplementary information. In the areas of psychophysics and scaling, where references are scattered and reviews are few, Guilford's treatment is more thorough than in the areas of testing and factor analysis, where other good summaries are available. There was obviously not space enough to include all of the new techniques and ideas. The up-and-down method, probit analysis, and Coombs' general approach to scaling all deserve more extended treatment than they receive. In general, though, Guilford's coverage is excellent.

Guilford's treatment of the method of successive categories is the weakest section in the book. In contrast to the usual clarity of exposition, this section is fuzzy and very difficult to follow. (There are some printing errors to add to the confusion). Some of the details are right, others are wrong, or at least dubious. His method for estimating category boundaries—or limens—is standard, but then he suggests locating the stimuli by finding the interpolated medians of the judgment distributions on this scale of limens. According to him, means are harder to find, and in either case, trouble arises when judgment distributions are truncated, i.e., when many judgments are in an extreme category. Actually, when appropriate procedures are used, the stimulus means are easy to determine, and the method is indifferent to truncation. The basic difficulty with the presentation is that the successive categories model is never stated explicitly. In fact, Guilford seems to reject the model when he argues that the categories themselves should somehow be scaled rather than the boundaries between categories. His procedure for scaling the categories makes no sense to this reviewer. Since the method of successive categories has great utility, this section of the book is especially disappointing.

Guilford's exposition is usually very clear, and his style is straightforward. The clarity is slightly compromised by the fact that Σ never appears with an index of summation or limits. This is sometimes confusing, although seldom ambiguous. The text includes many examples that help the reader to follow the development. However, it would have been better strategy to draw more psychophysical examples from sensory psychology. Emphasis on problems like discrimination, masking, and target detection in vision and audition, rather than on lifted weights might interest students who now find psychophysics dull.

In a book of this sort, it is sometimes necessary to introduce formulas magically, either because there is not space to derive the formulas, or because the development would be beyond the mathematical abilities of most students. In *Psychometric Methods* magic is used quite frequently. There are several places where a few words of explanation would reveal the trick and allow the student to understand the development or at least to get an intuitive grasp of the idea. Some examples will be cited here. In describing scoring formulas for tests, Guilford states that $R - W/(k - 1)$ is perfectly correlated with $R + B/k$. If the student realizes that rights plus wrongs plus blanks equals the total number of items on the test, he can easily prove the assertion: the insight should have been supplied in the text. Fitting a straight line by the method of averages is very simple to comprehend if the student realizes that he is really selecting two subgroups of points and drawing a straight line through the means of these subgroups. However, Guilford's purely algebraic discussion is likely to seem magical.

Another example is provided by Tucker's version of Kuder-Richardson formula 20, which is written so that a priori estimates of the variance of the item p 's can be inserted. It is not stated in the text that Tucker's formula is algebraically equivalent to K. R. 20. Indeed, the text implies a lack of equivalence. A slightly different case is the formula for estimating discriminial dispersions in Case III of the law of comparative judgment, which is presented uncritically. The inquisitive student will discover that the approximation is based on some very tenuous assumptions and may not be very approximate. The student should have been warned that this one is done with mirrors.

The book has its share of errors. Most of these are in the numerical examples, the table headings, and in the formulas. Three are worth noting here. In formula (10.7), p. 253, the S^2 should be preceded by Σ . In formula (14.17), p. 386, the Σ in the denominator should be inside the parentheses. In formula (16.1), p. 472, the radicals indicating the fourth root of the numerator and the square root of the denominator are omitted. Most of the other errors that we noticed will not bother an alert self-confident student.

The appendix contains a fine collection of useful tables. Table C, which gives normal

deviates and ordinates for various values of the area, is especially valuable and can be found almost nowhere else.

The major changes in *Psychometric Methods* are in scope and organization. The reader's evaluation of the second edition can be predicted accurately from his estimate of its predecessor.

Massachusetts Institute of Technology

Bert F. Green

C. RADHAKRISHNA RAO. *Advanced Statistical Methods in Biometric Research*. New York: John Wiley and Sons, 1952, pp. xvii + 390.

This book should be of much interest to social scientists and other investigators who are so often confronted with data requiring multivariate analysis of one kind or another. The style, which presupposes a working knowledge of elementary statistics, is a combination of terse mathematical statement followed by examples, mostly from the fields of anthropometry and genetics. A psychological application (p. 316, p. 370) is of importance, for it illustrates the solution to the problem of "types," under the restriction that measurement dispersion matrices for the types are identical.

The first chapter neatly summarizes that part of matrix algebra most useful in statistics, including quadratic forms. Also, the technique of pivotal condensation for evaluating determinants and matrix inverses is first discussed here, and throughout the book the value of this method for simplifying computations is ably demonstrated.

The second chapter gives statistical distributions in common usage, followed by the multivariate distributions required for tests treated in subsequent chapters. Some practical insight into the use of distributions for constructing multivariate tests is provided by this chapter and Chapter 7.

The remaining chapters are oriented toward testing hypotheses, with adequate emphasis on cases where variates are correlated. The last three chapters contain, with only minor revisions, the author's previously published contributions to the theory and use of classification, or discrimination, functions. The value of this work for psychologists and anthropologists can hardly be overemphasized.

Chapter 4 contains an interesting and original presentation of maximum likelihood estimation, where the Fisherian concepts of efficient scoring and amount of information in scores are illustrated. Chapters 3-6 contain useful sections on many of the traditional problems of inferential statistics. Analysis of variance is discussed only briefly, but a new technique for obtaining an interaction sum of squares is given, and a problem requiring classical analysis of covariance is fully illustrated. (Generalized analysis of variance and covariance, or "analysis of dispersion," is treated in Chapter 7.). The sections on chi-square are clear and relatively complete; for example, included is the evaluation of 2×2 tables with more than one degree of freedom, the use of Dandekar's (instead of Yates') correction, and a more exact approximation to the normal distribution than that obtained by using $\sqrt{2\chi^2} - \sqrt{2n - 1}$. An equation on page 197 is incorrect, but in the context the slip is obvious to the reader.

Several appendices are included, one of which contains a number of original lemmas on classificatory problems; another contains two methods for applying a Schmidt transformation to obtain uncorrelated variates.

The main disadvantage of the book lies in the fact that most readers who will want to use the methods may find it difficult to make rather abrupt transitions from very general mathematical thinking to concrete applications. In other words, the book may be too difficult for those for whom the applications seem most pertinent. Also, psychologists may

be disappointed that the final chapter contains so little on factor analysis. The author makes use of canonical variates and correlations without clearly relating them to Hotelling's method of factor analysis. But such criticisms are unimportant in view of the many remarkable contributions so adequately and creatively utilized in this volume.

Mellon Foundation
Vassar College

Harold Webster

RAYMOND B. CATTELL. *Factor Analysis: An Introduction and Manual for the Psychologist and Social Scientist*. New York: Harper & Bros., 1952. Pp. xiii + 462.

In the Preface of *Factor Analysis*, Cattell has set forth three principal requirements which the book should fulfill. (1) "—to meet the need of the general student in science to gain ideas of what factor analysis is about and to understand how it integrates with scientific methods and concepts generally," (2) to serve "as a textbook for statistics courses which deal with factor analysis for the first time, either as an appreciable part or as the whole of the semester course," and (3) "—to supply a handbook for the research worker, the student, and the statistical clerk which will be a practical guide with respect to carrying out the processes most frequently in use." To achieve these three objectives Cattell has written the book in three sections: I Basic Concepts in Factor Analysis, II Specific Aims and Working Methods, and III General Principles and Problems. Each of these sections has been planned in terms of a sound psychological approach to teaching in which the reader is carried from simple concepts to more complex ones, from general principles to specific items, and from elementary numerical examples to illustrative problems involving numerous intricate steps.

It is helpful to evaluate the contribution of the book in terms of the extent to which it appears that the author has attained each of the three requirements set forth. Although giving the impression of being somewhat missionary in his remarks concerning the scope of the applications of factor analysis, Cattell has done well in explaining at a readily-grasped intuitive level the basic principles underlying factor analysis and in stating the numerous uses to which the factor analytic techniques can be put. One of the strongest features of the text is the thorough and penetrating discussion of the place of factor analysis in the design of experiments. In addition to explaining at length the characteristics and application of the R and Q , P and O , and S and T techniques, the author has attempted to relate these six procedures to features of the classical experimental design and to modern approaches involving use of analysis of variance in factorial designs. Moreover, he has shown explicitly the potentialities of factor analysis not only in theory construction, but also in applied fields of educational and social psychology. Not the least interesting of his achievements is the discussion of the nature of the interpretation of factors that appear in an analysis. In short, the first objective has been achieved in a noteworthy fashion.

Since the realization of the second objective concerning the textbook function actually depends upon the fulfillment of the third objective relating to the handbook service of the book, it is advantageous next to evaluate the degree to which the third objective has been attained. As a manual numerous shortcomings are apparent:

(1) It would appear that an attempt has been made to explain too many methods of centroid extraction, communality estimation, and factor extraction relative to the limited space devoted to those topics. A somewhat more extensive explanation of a fewer number of techniques might have been desirable.

(2) The steps involved in the various clustering methods do not seem to be easy for

the beginner to grasp, since the illustrative examples are not clearly related to the procedures described. For example, the explanation of the group method of factoring (pp. 174-8) seems to be unnecessarily confusing and ambiguous. The origin of the entries appearing in the table at the top of page 176 remains a mystery to the reviewer.

(3) The format of the computational explanations is such that one cannot grasp in a readily-apparent fashion the objectives toward which the writer is trying to lead the reader. Paragraph captions or headings would be particularly helpful. In short, each of the steps involved in the calculations is simply not clearly set forth for the reader to perceive. Each rule or procedural item should be directly related to a specific numerical operation.

(4) The explanation concerning the rotation process through use of graphs is substantially inadequate if the text is to serve as a manual. What is seriously needed is a set of graphs to illustrate in a step-by-step fashion the solution of a representative problem involving between 10 and 20 test variables. In addition, a paragraph or two in which an explanation is given as to why each rotation was undertaken would be most helpful to the beginning student. Both orthogonal and oblique rotations should be considered at much greater length. Although mastery of the art of rotation requires extensive experience, a list of guiding principles that are related to illustrative plots would constitute an important teaching aid.

(5) The presence of numerous errors is particularly annoying and confusing to both the beginner and the experienced worker. One rather serious mistake occurs in the equation near the top of page 232. Instead of F_n or $V_F = V_0 \lambda F$ the equation should be written as F_n or $V_F = V_0 \lambda F^{-1}$. There is some doubt as to whether the geometric interpretation of reflection in centroid extraction that is presented on page 54 is correct. Numerous minor errors are present. A few examples may be cited: a double negative on page 157, line 9, which would not seem to be intended; one numerical entry of 0.37 in line 3 of the second paragraph on page 160 when the value of 0.38 is intended; misplacement of the decimal point of the numerical entry in the denominator of the fraction appearing at the bottom of page 160; an incorrect numerical entry (4.45 instead of 4.72) in the denominator of the fractions from which m is calculated on page 172; an apparently erroneous value of .10 instead of about .15 in the second row and first column of Table 26 on page 201; the use of communality when square root of communality is intended on line 17 of page 205; the use of "are" when "is" is intended on the fifth line from the bottom of page 256; an incorrect reference to Table 27 on page 214, and least important the misspelling of the reviewer's name.

(6) Much needed is a summary in one location of the matrix equations that are frequently employed in factor-analysis studies—a set of 12 or 15 equations that show various interrelationships among the primary-factor, reference-factor, arbitrary-orthogonal-factor loadings, the intercorrelations of the factors (both types), and the relationship of correlation coefficients to various types of factor loadings. Such a summary would serve to unify much of the illustrative material.

These remarks represent to a large extent a consensus based on the numerous statements of students who have used the book as a text and upon the comments of professors who have either required the book in courses or have attempted to use it as a manual in their own research. In short, the third requirement has not been realized.

Since the book as a manual is somewhat limited in the clarity of its exposition with respect to the use of numerical procedures, the second requirement concerning its function as a textbook has not been met to an adequate degree. It would appear that the instructor in factor analysis would need to require a second text to supplement the content of Cattell's *Factor Analysis* if skills in factoring are to be gained. Students have consistently reported that it has been necessary to consult other sources at length to clarify what are essentially routine steps involved in clerical procedures.

One of the most pleasing features of the book is Cattell's style of writing, which is informal and conversational in its tone. His ample use of cleverly devised figures of speech such as similes, personifications, and metaphors offers many an opportunity for a smile as well as a refreshing change of perspective in the reader's orientation to the field of abstractions that pour forth page after page. A few examples may be definitive:

"This business of reflecting, however, can become as exasperating as trying to hold three footballs in two hands; for as we make r 's positive as a whole for one variable, we make some individual r 's in the column negative for other tests (p. 55).

"The search for common characteristics in the loaded variables which would give a first hunch as to the nature of the factor is beset by difficulties when the loadings are not very high, and always presents possibilities of being misleading. To take a trivial, not to say frivolous, example, if two drunken men and two sober men constituted our population and one of the former had had Scotch and soda while the other had had Bourbon and soda, but the sober men had had nothing we should obtain correlations suggesting a cluster or factor in which drunkenness and soda would be most strongly loaded. Only a person who knew that the variables—Bourbon and Scotch—contained the common influence alcohol would recognize the role of alcohol in the drunkenness syndrome; and only the choice of a sufficiently varied population to include persons who had drunk soda but not alcohol would reduce the soda variable to its proper negligible loading in the drunkenness factor (pp. 75-76).

"—and it has frequently happened that a reference vector which has obstinately eluded stabilization has been led to a recognizable hyperplane by this method as soon as all its fellow reference vectors have become sufficiently convincing in their hyperplanes to apply it (p. 213).

"The points are gradually being tracked down by these successive moves and shepherded into a restricted area as are sheep by a well-trained sheep dog" (p. 54).

In its current form the book is an excellent source for the person interested in the general principles of factor analysis, in the place of factor analysis in experimental design, and in types of problems in the social sciences for which factor analysis may be useful. However, as a guide or manual to be employed in the actual performance of a factor analysis the book is of doubtful value. Despite the limitations as a manual, it would be a useful supplementary text in beginning courses in factor analysis.

University of Southern California

William B. Michael

Kit of Selected Tests for Reference Aptitude and Achievement Factors. Educational Testing Service, Princeton, New Jersey: October, 1954.

This kit contains three reference tests for each of sixteen aptitude and achievement factors, as well as a manual giving detailed information about these tests. The purpose of the kit is stated as follows by French: "Tests in this kit are suggested for use in factorial studies where representation is desired for any of the . . . aptitude or achievement factors (included) . . . It is intended that use of the Kit tests for defining reference factors will facilitate interpretation and the confident comparison of one factor study with another." The factors to be included were chosen and the tests selected by a variety of overlapping committees, whose work was mainly done by correspondence. Tests and manual are contained in a strong folder which should help in keeping them together in one place.

While the work of assembly was obviously well done, a number of disturbing thoughts will occur to the reader. Is the democratic process of committee discussion really well adapted to the production of scientific truth? Would physicists consent with equanimity

to have elements defined in terms of voting and a show of hands rather than in terms of decisive proof? Would the inclusion of non-American writers (Vernon, Meili) have decisively altered the contents of this kit? These are important problems, but they are not discussed in the manual. Perhaps the employment of some such technique as Ahmavaara's (Ahmavaara, Y. Transformation analysis of factorial data. Helsinki: Suomalaisen Tiedeakatemian. 1954) might have helped to objectify judgments.

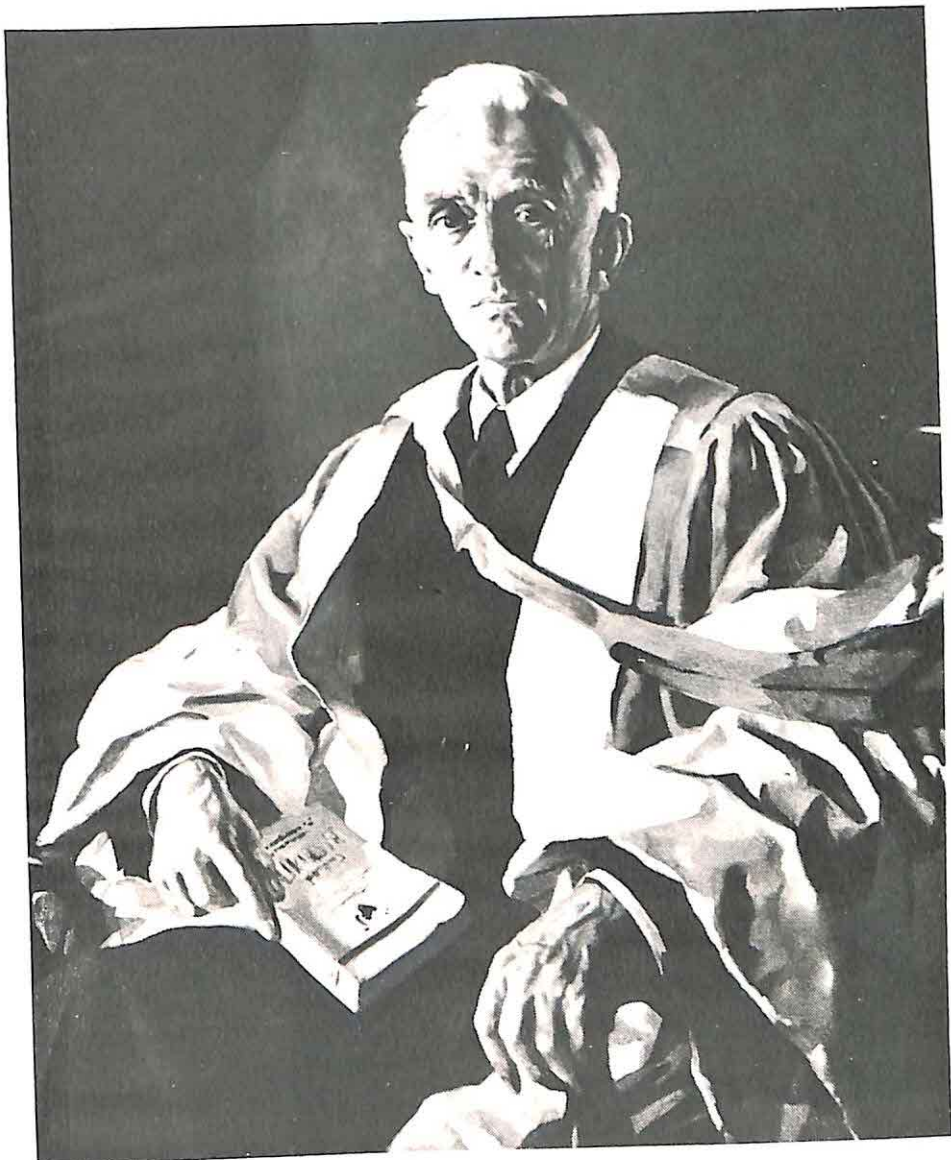
Another difficulty will strike many readers. Some of the factors possess a high degree of generality, such as the general reasoning factor; others, like the aiming factor, are very specific indeed. To put both types, as well as factors of intermediate coverage, into the same kit raises problems of specificity and generality which, again, are not discussed in the manual. Nor is there a discussion of the very meaning of the term "aptitude" used to designate these different types of factors. In the reviewer's department, tests with high loadings on the "aiming" factor have been found to be excellent measures of temperamental abnormality; to what extent can we rest content with having them treated as pure ability tests?

Another difficulty that arises is due to the failure of the committees to consider evidence from outside the factorial field. To take but two simple examples, we may wonder to what extent reactive inhibition (I_R) and conditional inhibition (${}_sI_R$) play a part in the performance of dotting, aiming, and tracing tests. To what extent, also, do individual differences in inhibition formation determine score? On a rather higher plane we may ask about the determination of the results on all the tests included of orectic factors, which have been shown (Furneaux, W. D. Some speed, error, and difficulty relationships within a problem-solving situation. *Nature*, 1952, 170, 37) to exert an important and variable influence. The fact that none of these objections are discussed or met by French is less his fault than that of factor analysts, who in general tend to pay little attention to the findings of general psychology in their work. Nevertheless, it does tend to make this collection less valuable than it might otherwise have been. It also gives a false feeling of security to investigators who may wish to work in this field.

This brings us to the last point. The kit apparently is intended for research workers who may wish to use these tests as reference markers. It is difficult to see why in this case the actual tests have been included with the manual. Research workers in any case will have to obtain sets of tests which they wanted to employ, and they would certainly be expected to be familiar with the current literature and the tests included in the kit if the research to be done were to be taken seriously. For the purpose of ensuring the use of reference markers, therefore, the manual itself would have been quite sufficient. It seems to the reviewer that the main use of the kit will be, not for research workers, but for instructors who wish to show their students illustrative tests of the main factors isolated by factor analysts. For this purpose the kit is admirably selected and constructed; it is to be hoped, however, that in his discussion the instructor will not forget to include some of the matters raised in the first few paragraphs of this review.

*Department of Psychology
Institute of Psychiatry
University of London*

H. J. Eysenck



SIR GODFREY THOMSON

Sir Godfrey Thomson

It was at the International Congress of Psychology, 1923, that I first met Godfrey Thomson. We were in the same symposium on the nature of intelligence. In correspondence and in personal conferences I have found him always friendly and intellectually generous even when we did not agree in our psychological interpretations. I always read his criticisms with interest and respect. An outstanding characteristic was that he never falsified a problem in order to win an argument—a trait that was not shared by some of his adversaries in the controversies of mental measurement.

Godfrey Thomson was born in Carlisle, England, on March 27, 1881. He was educated at Rutherford College, Armstrong College (now King's College), the University of Durham, and the University of Strasbourg. At Armstrong College he was Open Exhibitioner, Junior Pemberton Scholar, and Charles Mather Scholar. Later he was appointed Pemberton Fellow of the University of Durham, where he obtained the M.Sc. degree in mathematics and physics. Following this he attended the University of Strasbourg in Germany and was awarded the Ph.D., *summa cum laude*, in 1906.

At this point his interest turned from the physical sciences to psychology and he returned to the University of Durham for postgraduate study in that subject. After receiving the D.Sc. degree in Psychology in 1913 he accepted the position of Lecturer in Education at Armstrong College. In 1920 he became Professor and Head of the Department of Education; he held this position until 1925. During this period he visited the United States as Visiting Professor of Education at Columbia University, 1923–24. A second visit to this country was in 1933 when he was a lecturer in the Yale Summer School.

From 1925 until his retirement in 1951 he held the joint post of Professor of Education at the University of Edinburgh and Director of Studies, Edinburgh Provincial Committee for the Training of Teachers. In 1939 the University of Durham awarded him an Honorary D.C.L. Later he was awarded the Order of Polonia Restituta (third class) by the Government of Poland in exile, and in 1949 he was knighted. Sir Godfrey Thomson died in Edinburgh on February 9, 1955, at the age of 73.

Godfrey Thomson was a fellow of the Royal Society of Edinburgh, of the Eugenics Society, and of the British Psychological Society, of which he was president, 1945–46. He was an Honorary Fellow of the Educational Institute of Scotland, and of the Swedish Psychological Society. He was a member of the British Association for the Advancement of Science, the National Institute of Industrial Psychology, the International Statistical Institute, and of a large number of boards and foundations.

Sir Godfrey Thomson had many connections with scientific societies in the United States: Foreign Honorary Member of the American Academy of

Arts and Sciences, Foreign Associate of the United States National Academy of Sciences, Fellow of the American Association for the Advancement of Science, member of the American Institute of Mathematical Statistics, and member of the Psychometric Society.

He devised tests of intelligence and achievement which were well-known and widely used in the British Isles and throughout the Commonwealth. With the profits from the sale of these tests he founded scholarships and endowed the Godfrey Thomson Lectureship in Educational Research in Edinburgh University.

Sir Godfrey Thomson's work in mental measurement can be divided into three successive periods. First he was interested in psychophysical problems, beginning in 1911. His work was published in *Essentials of Mental Measurement* by Brown and Thomson and in a number of papers. The second period represents his work on the social and geographical distribution of intelligence and the influence of differential birth rate. A third period was devoted to the factorial analysis of human ability, a field which interested him the most. His work in this field is represented by his well-known book *The Factorial Analysis of Human Ability*, which has appeared in several editions. He described his main objective as an attempt "to bring mathematical exactitude into psychological experiment and theorizing."

Psychometric Laboratory
University of North Carolina

L. L. Thurstone

A GENERALIZED SIMPLEX FOR FACTOR ANALYSIS*

LOUIS GUTTMAN

THE ISRAEL INSTITUTE OF APPLIED SOCIAL RESEARCH,
JERUSALEM, ISRAEL

By a simplex is meant a set of statistical variables whose interrelations reveal a simple order pattern. For the case of quantitative variables, an order model was analyzed previously which allowed only for positive correlations among the variables and a limited type of gradient among the correlation coefficients. The present paper analyzes a more general model and shows how it is more appropriate to empirical data. Among the novel features emerging from the analysis are: (a) the "factoring" implied of the correlation matrix; (b) the use of a non-Euclidean distance function; and (c) the possible underlying psychological theories.

I. Introduction

In a new approach to factor analysis, called *radex* theory, it has been shown (3, 4) how two important special cases arise: the *simplex* and the *circumplex*. Only a restricted case of the simplex was considered parametrically in (3), allowing only positive correlations among the observed variables and only a limited type of gradient among the correlation coefficients. The purpose of the present paper is to give a parametric theory and analysis of a more general type of simplex. In this generalization, a more flexible gradient is possible, and negative correlations can appear as well as positive ones. Thus, "inhibiting" as well as "reinforcing" factors can be considered. Generalizing the parametric system for a simplex immediately suggests analogous generalizations for the circumplex, and hence also for a complete radex. We shall consider here only the simplex, and it will be clear what the implications are for the circumplex and radex.

As in conventional factor analysis, we consider a universe of tests for a population of subjects. Both the universe and the population are usually theoretically indefinitely large, and in practice only a finite sample is drawn from each. It will be convenient to consider a finite battery of n tests from the universe, but to consider the population of testees to be infinitely large so that we need not be concerned with sampling error due to people. We shall then be able to see what happens as n increases.

A particularly curious result of the present analysis is as follows. It turns out that in terms of ordinary factor analysis, one should factor not the co-

*Read at the International Congress of Psychology, Montreal, June 7-12, 1954. This research was facilitated in part by an uncommitted grant-in-aid to the writer from the Behavioral Sciences Division of the Ford Foundation.

variance matrix of our generalized simplex, but rather the *inverse* of this matrix. The factoring implied is of two kinds. First, the *first centroid*—in the sense of Thurstone—should be factored from the *inverse* matrix, and then the *principal components* should be taken as the remaining $n - 1$ factors. This particular way of regarding the factor resolution turns out to have important theoretical and practical implications for the present simplex theory.

A second and most highly important result reveals a limitation of considering variables as points only in a Euclidean space. Regarded this way, our simplex appears n -dimensional, or with as many Euclidean dimensions as distinct variables. However, when distances between these same points are measured in a certain non-Euclidean fashion, then the points can be plotted on a straight line, or they form a *one-dimensional* non-Euclidean system.

Further novel features appear in our generalized simplex with respect to the psychological theories that can possibly account for it.

II. General Notation

Let t_{ij} denote the observed score of person i on test j . The mean and the standard deviation of each test are arbitrary, and indeed are usually artifacts of the test construction procedure (3). One part of the problem of factor analysis is to express each t_{ij} as the sum of two types of components: common and deviant (or "unique"). Let e_{ij} be the score of person i on the deviant component of test j . Then we can write, for all i and j ,

$$t_{ij} = w_i s_{ij} + e_{ij}, \quad (1)$$

where s_{ij} is the structural or non-deviant part of t_{ij} , and w_i is a multiplying constant to allow for the arbitrariness of the standard deviation of the observed t_{ij} . Especially in the simplex theory to follow, the standard deviations of the s_{ij} are *not* in general arbitrary.

Since the present simplex theory is concerned only with covariances between the s_{ij} , it will be convenient to consider the mean of each to be zero,

$$E_i s_{ij} = 0 \quad (j = 1, 2, \dots, n). \quad (2)$$

Various laws of deviation are possible for the e_{ij} , as pointed out in (3). The one assumed in conventional factor analysis is the δ -law,

$$\text{cov}(e_i, s_k) = \text{cov}(e_i, e_k) = 0 \quad (j \neq k). \quad (3)$$

A well-known consequence of (3) and (1) is that

$$\text{cov}(t_i, t_k) = w_i w_k \text{cov}(s_i, s_k) \quad (j \neq k). \quad (4)$$

According to (4), the covariance matrix of the observed tests is derived from that of the underlying s_i merely by constants of proportionality, except

for the main diagonal. Any submatrix in the one that involves no main diagonal element must have exactly the same rank as the corresponding submatrix in the other. This suggests one way of testing hypotheses about the s_{ij} , insofar as these lead to conditions on the ranks of certain submatrices.

The δ -law (3) may or may not be true in practice for a given set of data. One approach to testing it for some data is by image analysis (6). We shall be concerned here primarily with *structural* laws or theories for the s_{ij} , and the truth or falsity of the deviance law (3) is a subsequent problem to be explored ultimately with empirical data in any given case.

III. Review of Previous Data and Theory

Several correlation matrices published earlier in the literature by various writers have now been re-analyzed and found to form approximate simplexes. These data represent a wide variety of mental abilities and personality traits (1, 3, 4). Two examples are shown in Tables 1 and 2. One is of a battery

TABLE 1
Correlations Among Six Numerical Ability Tests*

Test	Addition	Subtraction	Multiplication	Division	Arithmetical Reasoning	Numerical Judgment
Addition	1.00	.62	.62	.54	.29	.28
Subtraction	.62	1.00	.67	.53	.38	.37
Multiplication	.62	.67	1.00	.62	.48	.52
Division	.54	.53	.62	1.00	.62	.57
Arithmetical Reasoning	.29	.38	.48	.62	1.00	.64
Numerical Judgment	.28	.37	.52	.57	.64	1.00

*From Table 2, pp. 110-112 of (11). See analysis in (3).

TABLE 2
Correlations Among Six Tests of a Certain Type of Verbal Ability*

Test	Proverbs	Vocabulary	Word Checking	Verbal Enumeration	Association	Synonyms
Proverbs	1.00	.55	.29	.24	.18	.17
Vocabulary	.55	1.00	.46	.44	.31	.24
Word Checking	.29	.46	1.00	.56	.34	.22
Verbal Enumeration	.24	.44	.56	1.00	.43	.27
Association	.18	.31	.34	.43	1.00	.45
Synonyms	.17	.24	.22	.27	.45	1.00

*Called "abstractness of verbalization" in (4, p. 13). Data from Appendix Table 1 of (12): tests 43, 45, 58, 57, 6 and 55.

of numerical ability tests, and the other is of a certain type of verbal ability tests.

From mere inspection of Tables 1 and 2, it is clear that there is some kind of order relationship within each battery of tests. In each case, the largest correlations are next to the main diagonal, and taper off to the north-east and southwest corners of the table. No other arrangement of the rows and columns of the tables, or reshuffling of the order of the variables, will yield such an apparent gradient. It is as if one could regard the variables to be points ordered along a straight line, and the correlation of one variable with another decreases as the other departs from it—in either direction—along this line.

One of the interesting new parametric properties to be developed in the simplex theory of the present paper is that simplex variables can be *literally* plotted as points along a straight line, with distances between them being strictly additive.

It has been shown in (3) that it is possible to write a factor model which will yield a gradient among correlation coefficients that has the general characteristics of the empirical ones in Tables 1 and 2 (or of the several other known empirical examples of approximate simplexes). For example, assume there are n uncorrelated factors underlying the n tests in the battery. Let x_{ic} denote the score of person i on factor x_c . It is convenient to assume also that the means of the x_{ic} are zero. Thus, the assumptions so far can be written as

$$E_i x_{ic} = 0 \quad (c = 1, 2, \dots, n) \quad (5)$$

and

$$E_i x_{ib}x_{ic} = 0 \quad (b \neq c; b, c = 1, 2, \dots, n). \quad (6)$$

Now, assume further that there is an order within the s_i and also within the x_c such that for all i and j the following factor law of formation holds:

$$s_{ij} = \sum_{c=1}^j x_{ic} \quad \left[\begin{array}{l} \text{Additive} \\ \text{restricted} \\ \text{simplex} \end{array} \right]. \quad (7)$$

Let σ_{s_i} and σ_{x_c} be the standard deviations of s_i and x_c , respectively, and let $\rho_{s_i s_k}$ be the coefficient of correlation between s_i and s_k . Then it has been proved from (5), (6), and (7) that

$$\sigma_{s_j}^2 = \sum_{c=1}^j \sigma_{x_c}^2 \quad (j = 1, 2, \dots, n) \quad (8)$$

and

$$\rho_{s_i s_k} = \sigma_{s_i} / \sigma_{s_k} \quad (j \leq k). \quad (9)$$

According to (8), σ_{s_k} increases as k increases, so that for fixed j in (9) the right member must decrease as k departs from j . This describes a gradient in the correlation coefficients of the s_j , which when modified in the t_i by the presence of error as in (1)—say that (3) and (4) hold—can approximately give rise to observed gradients such as in Tables 1 and 2.

Another way of writing law (7) is

$$s_{ij} = s_{i,j-1} + x_{ij} . \quad (10)$$

Equality (10) asserts that s_j is the same as its predecessor s_{j-1} , except for the addition of a new factor. Interpreting Table 1 this way would imply that—apart from deviant factors of the e_j type—the subtraction test involves the same x_1 as does the addition test, but also an x_2 not called on by the addition test. The multiplication test calls on both x_1 and x_2 , but also on an x_3 , etc. A corresponding explanation would hold for the hierarchy among the verbal ability tests of Table 2.

It has been shown in (3) how an entirely different factor law can give rise to exactly the same type of correlation matrix as in (9). Instead of having factors x_e that are added according to (7), it is possible to write a law wherein factors are *multiplied* by each other and yet yield a hierarchy of correlations *identical* with (9). Even other laws may yield exactly the same results.

But it has also been pointed out in (3) that the detection and use of the simplex pattern does not at all depend on knowing whether law (7) holds or some alternative law leading to identical results. It is sufficient to determine the law of formation of the *correlation coefficients*, say such as (9), and for this the specification of an underlying law of factors such as (7) is not strictly necessary.

An important feature of a matrix with elements of the form (9) is that, if $\sigma_{s_j} \neq \sigma_{s_k}$ whenever $j \neq k$, then the matrix is nonsingular. Furthermore, the inverse of this nonsingular matrix has zero elements everywhere except in the main diagonal and in the immediately adjacent diagonals. This has profound implications for prediction problems, since the elements of the inverse matrix are the basis for the multiple regression weights for any linear multiple regression on the s_j . This also has profound implications for the internal structure of the s_j , for these vanishing elements of the inverse show that the principal components of the s_j satisfy a certain second-order linear difference equation, and hence must obey a certain general oscillatory law of formation (2, 3).

We now wish to generalize law (9). We shall do this in two steps. The first stage is to use a generalization of law (7) for expository purposes.

IV. A First Parametric Generalization of the Additive Simplex

It is clear from (9) that only positive correlations can arise from the restricted hypothesis (7). But surely there must be an order system which

would also allow for negative correlations. It is also verifiable from (9) that any tetrad, or second-order minor determinant, must vanish if all of its elements are on one side of the main diagonal of the correlation matrix (and *not* vanish if elements come from both sides of the main diagonal). Could there be an order system that does not lead to such a restrictive condition on the rank of parts of the matrix?

A generalization of (7) that does relax these restrictions somewhat is as follows. In (7), each x_c operates as an "all or none" affair, in the sense that s_i does not involve x_c whenever $c > j$. For $c > j$, then, let us assume there is an *alternative* set of factors operating, say some y_c .

Let y_{ic} be the score of person i on alternative factor y_c . For convenience, assume the means of the y_c are zero

$$E_i y_{ic} = 0 \quad (c = 1, 2, \dots, n). \quad (11)$$

Analogous to (6), we assume the y_c to be uncorrelated with each other,

$$E_i y_{ib} y_{ic} = 0 \quad (b \neq c; b, c = 1, 2, \dots, n). \quad (12)$$

We also assume y_c to be uncorrelated with x_b whenever $b \neq c$,

$$E_i x_{ib} y_{ic} = 0 \quad (b \neq c). \quad (13)$$

Let γ_c denote the covariance between x_c and y_c

$$\gamma_c = E_i x_{ic} y_{ic} \quad (c = 1, 2, \dots, n). \quad (14)$$

No assumptions will be made here concerning the size or sign of γ_c for any c ($c = 1, 2, \dots, n$). Different covariances can arise from different psychological processes. For example, if x_c is an "excitatory" factor, then y_c might be an "inhibiting" factor, and the covariance γ_c might be negative. Or x_c and y_c might denote two different levels of excitation (or of inhibition) of the same type of factor, and hence γ_c might be positive.

We can now write the following generalization of law (7)

$$s_{ij} = \sum_{c=1}^j x_{ic} + \sum_{c=j+1}^n y_{ic} \quad \left(\begin{array}{l} \text{First generalization} \\ \text{of additive simplex} \end{array} \right). \quad (15)$$

In place of (8) we now get

$$\sigma_{s_j}^2 = \sum_{c=1}^j \sigma_{x_c}^2 + \sum_{c=j+1}^n \sigma_{y_c}^2 \quad (j = 1, 2, \dots, n). \quad (16)$$

It is also easy to derive from (11), (12), (13), (14), and (15) that

$$\text{cov}(s_j, s_k) = \sum_{c=1}^j \sigma_{x_c}^2 + \sum_{c=k+1}^n \sigma_{y_c}^2 + \sum_{c=j+1}^k \gamma_c \quad (j \leq k). \quad (17)$$

From (17) and (16)

$$\text{cov}(s_j, s_k) = \sigma_{s_j}^2 + \sum_{c=j+1}^k (\gamma_c - \sigma_{v_c}^2) \quad (j \leq k), \quad (18)$$

so that (9) generalizes to

$$\rho_{s_j s_k} = (\sigma_{s_j} / \sigma_{s_k}) + \left[\sum_{c=j+1}^k (\gamma_c - \sigma_{v_c}^2) \right] / (\sigma_{s_j} \sigma_{s_k}) \quad (j \leq k). \quad (19)$$

Since the second terms on the right of (18) and of (19) can be negative—especially when $\gamma_c < 0$ for some or all of the c —the left members can also be negative upon occasion. Thus, law (15) allows also for possible negative correlations among the s_{ij} .

The rank condition on the correlation matrix resulting from (9) is also relaxed a bit, according to (19). To see this, it is easiest first to deal with the covariance matrix defined by (18). Taking first differences with respect to k , we see that

$$\text{cov}(s_j, s_{k+1}) - \text{cov}(s_j, s_k) = \gamma_{k+1} - \sigma_{v_{k+1}}^2 \quad (j \leq k). \quad (20)$$

In the matrix of order $n \times (n - 1)$ defined by the left member of (20), all submatrices with elements all on one side of the main diagonal are clearly of rank one at most, according to the right member of (20). Hence, in the $n \times n$ matrix of the elements defined by (18), all corresponding submatrices cannot be of rank greater than two. But the rank of any submatrix in $[\rho_{s_j, s_k}]$ is the same as of the corresponding submatrix in $[\text{cov}(s_j, s_k)]$ since the rows and columns of one differ from those of the other only by constants of proportionality. Hence the rank of any submatrix of $[\rho_{s_j, s_k}]$ cannot be greater than two when all its elements are on one side of the main diagonal.

Formula (19) will of course allow for a closer fit to data such as in Table 1 and Table 2 than will formula (9). This may be needed especially to account for the aberration of the subtraction test from a simple gradient; apparently subtraction differs from addition and multiplication in somewhat of another manner than called for by law (7), and law (15) may be more appropriate.

V. A Second Parametric Generalization

A formulation like (15) is helpful in trying to understand what kinds of processes can possibly give rise to order relations among observed correlation coefficients. However, a formula like (19) can divert attention from the main consequences of having order relationships. One might be tempted to focus, for example, on the problem of estimating the γ_c and the $\sigma_{v_c}^2$ to be used in (19). Clearly, an analysis based on observed correlation coefficients alone can only hope at best to estimate the *differences* $(\gamma_c - \sigma_{v_c}^2)$, and not each term separately. That is, a correlational analysis alone cannot hope to piece out all the details of a process such as (15). Even if this were possible,

there are many important things to be learned about $[\rho_{s_i, s_k}]$ that do not need specification of these details.

We shall now give the main generalization of the simplex intended in this paper. It involves no explicit use of underlying factors x_c , y_c , or any others. Its focus is on what can be learned by a correlational analysis alone.

Each of laws (7) and (15)—given also the assumptions (6), (12), and (13)—satisfies the following necessary condition

$$E_i (s_{ij} - s_{ik})(s_{ik} - s_{il}) = 0 \quad (j \leq k \leq l). \quad (21)$$

This is an order condition among the s_i , and yet needs no detailed specification of an underlying factor mechanism. All that is hypothesized in (21) is that the difference between an s_k and any of its *predecessors* in the sequence is uncorrelated with the difference between this same s_k and any of its *successors* in the sequence: $s_k - s_j$ is uncorrelated with $s_k - s_l$ whenever $j \leq k \leq l$.

An interesting immediate consequence of (21) is that we can regard the s_i not merely as points arranged in a rank order, but we can specify an *additive metric* for distances between these points. Let d_{jk} be defined by

$$d_{jk} = E_i (s_{ij} - s_{ik})^2 \quad (j, k = 1, 2, \dots, n). \quad (22)$$

Now, we can write the identity

$$s_{ij} - s_{il} \equiv (s_{ij} - s_{ik}) + (s_{ik} - s_{il}). \quad (23)$$

Taking expectations of the squares of both sides of (23) shows that the following theorem is true.

THEOREM 1. *A necessary and sufficient condition for the order relation (21) to hold is that*

$$d_{il} = d_{jk} + d_{kl} \quad (j \leq k \leq l), \quad (24)$$

where d_{jk} is defined as in (22).

Therefore, if we define d_{jk} to be the *distance* between points s_j and s_k , this distance function is *additive* according to (24). If s_k is between s_j and s_l , then the distance from s_j to s_l is the sum of the distances from s_j to s_k and from s_k to s_l . Accordingly, the n points s_i can be plotted on a straight line, with distances between each pair being determined by formula (22).

It has been customary in factor analysis to regard all variables involved as being in a *Euclidean* space. For such a space, the distance between two points s_j and s_k is defined as the *square root* of d_{jk} . This makes the dimensionality of the space necessarily equal to the rank of the matrix $[\rho_{s_j, s_k}]$. Now this matrix is in general nonsingular when (21) holds, or n Euclidean dimensions are required. Using the non-Euclidean metric of (22) leads to but a one-dimensional space, according to Theorem 1.

It should be remarked that the distance function (22) does not yield a metric space in the general case of arbitrary variables, for the requisite triangular inequality need not be satisfied. However, we are using it here only for the special case where (21) holds, so the space of the specific points involved is certainly metric, being even one-dimensional in the sense of (24).

The writer first used a metric of the type (22) in the context of the principal components of scale analysis of qualitative data (8), and this suggested the developments presented here for a simplex of quantitative variables.

VI. The Rank of Certain Submatrices

From now on we shall be concerned largely with the covariances among the s_j , so it will be convenient to let σ_{jk} denote the covariance between s_j and s_k ,

$$\sigma_{jk} = \text{COV}(s_j, s_k) = \mathbb{E} s_{ij}s_{ik} \quad (j, k = 1, 2, \dots, n). \quad (25)$$

We wish to prove the following theorem:

THEOREM 2. *If n variables s_j satisfy the order condition (21), then any submatrix of $[\sigma_{jk}]$ cannot be of rank greater than 2 if all its elements are on one side of, or on, the main diagonal.*

For the proof, we first expand (21), using notation (25), to obtain

$$\sigma_{jk} = \sigma_k^2 + \sigma_{jl} - \sigma_{kl} \quad (j \leq k \leq l). \quad (26)$$

Since $[\sigma_{jk}]$ is a symmetric matrix, it suffices to consider only submatrices on one side of the main diagonal, say with all elements to the right of (or above) the diagonal. By differencing (26) with respect to j we see that

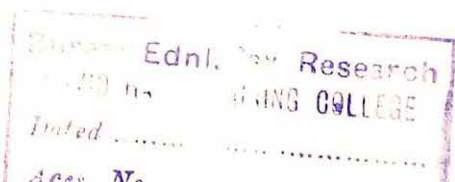
$$\sigma_{j+1,k} - \sigma_{jk} = \sigma_{j+1,l} - \sigma_{jl} \quad (j+1 \leq k \leq l). \quad (27)$$

According to (27), all elements to the right of the main diagonal and in the same row of the $n \times (n-1)$ matrix $[\sigma_{j+1,k} - \sigma_{jk}]$ are equal. Hence no submatrix which is all to one side of the main diagonal can have a rank exceeding unity. Consequently, the corresponding submatrices in $[\sigma_{jk}]$ cannot have ranks greater than 2, or Theorem 2 is proved.

VII. The Problem of Weights for Principal Components

Related to Theorem 2, but perhaps more striking, are two laws of formation: one for the inverse matrix and one for the principal components of $[\sigma_{jk}]$ when (21) holds.

In developing these laws, we first wish to take into account the fact that the principal components of a covariance matrix depend in part on the weight functions used, or the relative sizes of the standard deviations of the variables concerned. The components may shift also as one removes



variables from the matrix, or introduces additional ones. To see the effects of these operations when (21) is true, we shall introduce further notation.

First, we shall allow for the possibility that there is a frequency distribution over our n points s_i . This can arise from the fact that any observed variables t_k are but a sample from an infinite universe of variables. While each t_k has a different e_k , many can have exactly the same s_i , or be aimed at exactly the same aspect of the underlying simplex. Let f_i be the relative frequency of s_i in this sense; that is, f_i is the proportion of all the t_k which have the same s_i in (1). Then

$$\sum_{i=1}^n f_i = 1. \quad (28)$$

Next, we shall allow for the possibility that it is not the σ_{ik} themselves to be analyzed, but perhaps the $\rho_{s_i t_k}$, or some other weighted function of the σ_{ik} . Let v_i be the weight associated with s_i . Thus, if the principal components of $\rho_{s_i t_k}$ are to be analyzed, then $v_i = 1/\sigma_i$. If the principal components of the $t_i - e_i$ are to be analyzed, as in (1), then $v_i = w_i$. In general, the v_i represent any set of real numbers, and we wish to know the principal components of the Gramian matrix $[v_i v_k \sigma_{ik}]$ when relative frequency f_i is associated with row and column j .

Let λ denote a latent root of the matrix, and let z_i be the j th element of the associated latent vector. Our job is to solve the stationary equations (cf. 3)

$$\sum_{i=1}^n z_i f_i v_i v_k \sigma_{ik} = \lambda z_k \quad (k = 1, 2, \dots, n). \quad (29)$$

To simplify notation for the solution, let

$$u_i = z_i/v_i, \quad a_i = f_i v_i^2 \quad (j = 1, 2, \dots, n). \quad (30)$$

Then (29) can be rewritten as

$$\sum_{i=1}^n a_i u_i \sigma_{ik} = \lambda u_k \quad (k = 1, 2, \dots, n). \quad (31)$$

It should be remarked that, from (30), the a_i are always non-negative, even though the v_i may be negative. There is no loss of generality, then, assuming all the a_i to be positive,

$$a_i > 0 \quad (j = 1, 2, \dots, n), \quad (32)$$

for if $a_i = 0$, this would be equivalent to $f_i = 0$, or no s_i to begin with to use in (31).

VIII. Deriving the Inverse of the Covariance Matrix

It will prove convenient to study (31) by means of the inverse of $[\sigma_{ik}]$. This is more than just a matter of convenience, for the inverse matrix is of

basic importance in its own right. It provides the regression coefficients in multiple correlation problems involving all the s_i , and it provides the partial correlation and multiple correlation coefficients involved. In short, it is a basic tool of image analysis (6).

Eventually one would like to know about the inverse of the observed correlation matrix $[\rho_{i,j}]$. Since this will depend partly on the deviance law of the e_i in (1), all we shall do in the present paper is analyze the case where there is no error; we shall concentrate only on $[\sigma_{jk}]$. But even so, it is important to allow for the frequency function f_i , and to be concerned ultimately with the infinite universe of variables and not just a finite observed sample therefrom (6, 7).

If $[\sigma_{jk}]$ is nonsingular, let σ^{jk} denote the typical element of the inverse matrix. The inverse must be symmetric, since the covariances are. Thus,

$$\sigma^{jk} = \sigma^{kj} \quad (j, k = 1, 2, \dots, n). \quad (33)$$

If δ_{jk} denotes Kronecker's delta, then

$$\sum_{j=1}^n \sigma_{jk} \sigma^{jl} = \delta_{kl} \quad (k, l = 1, 2, \dots, n). \quad (34)$$

We shall solve (34) for σ^{il} by a differencing process.

The following differencing notation will be used. If z_k , z_{jk} , or z_{kl} are any quantities to be differenced with respect to k , then

$$\Delta_k z_k = z_{k+1} - z_k, \quad \Delta_k z_{jk} = z_{j,k+1} - z_{jk}, \quad \Delta_k z_{kl} = z_{k+1,l} - z_{kl}. \quad (35)$$

If we let $l = k + 1$ in (26), the equations can be rewritten as

$$\Delta_k \sigma_{jk} = \begin{cases} \sigma_{k,k+1} - \sigma_k^2 & (j \leq k) \\ \sigma_{k+1}^2 - \sigma_{k,k+1} & (j > k) \end{cases} \quad (k = 1, 2, \dots, n-1). \quad (36)$$

Differencing both members of (34) with respect to k and using (36) yield

$$(\sigma_{k,k+1} - \sigma_k^2) \sum_{j=1}^k \sigma^{jl} + (\sigma_{k+1}^2 - \sigma_{k,k+1}) \sum_{j=k+1}^n \sigma^{jl} = \Delta_k \delta_{kl} \quad \begin{pmatrix} k = 1, 2, \dots, n-1 \\ l = 1, 2, \dots, n \end{pmatrix}. \quad (37)$$

Let α_l be the sum of the elements in the l th column (row) of $[\sigma^{il}]$,

$$\alpha_l = \sum_{j=1}^n \sigma^{jl} \quad (l = 1, 2, \dots, n). \quad (38)$$

Also, notice from (22) that

$$d_{k,k+1} = \sigma_k^2 - 2\sigma_{k,k+1} + \sigma_{k+1}^2 \quad (k = 1, 2, \dots, n-1). \quad (39)$$

By bringing in the notion of a frequency function f_i we are in effect assuming our n points s_i to be distinct, or that

$$d_{k,k+1} > 0 \quad (k = 1, 2, \dots, n-1). \quad (40)$$

Let b_k and c_k be defined respectively as

$$b_k = (\sigma_{k+1}^2 - \sigma_{k,k+1}^2)/d_{k,k+1} \quad c_k = 1/d_{k,k+1} \quad (k = 1, 2, \dots, n-1). \quad (41)$$

Then by using (38), (39) and (41), we obtain from (37) that

$$\sum_{j=1}^k \sigma^{jl} = \alpha_l b_k - c_k \Delta_k \delta_{kl} \quad \begin{pmatrix} k = 1, 2, \dots, n-1 \\ l = 1, 2, \dots, n \end{pmatrix}. \quad (42)$$

Taking first differences in (42) with respect to k yields the important second-order difference equation

$$\sigma^{k+1,l} = \alpha_l \Delta_k b_k - \Delta_k (c_k \Delta_k \delta_{kl}) \quad \begin{pmatrix} k = 1, 2, \dots, n-2 \\ l = 1, 2, \dots, n \end{pmatrix}. \quad (43)$$

We now wish to obtain an explicit formula for α_l in (43). As is well known for Kronecker delta's,

$$\sum_{l=1}^n \delta_{kl} \equiv 1, \quad \sum_{l=1}^n \Delta_k \delta_{kl} \equiv 0. \quad (44)$$

Hence, if we let α be the sum of all n^2 elements of σ^{ik} , or

$$\alpha = \sum_{l=1}^n \alpha_l = \sum_{i=1}^n \sum_{k=1}^n \sigma^{ik}, \quad (45)$$

and if we sum both members of (42) over l , we obtain

$$\sum_{j=1}^k \alpha_j = \alpha b_k \quad (k = 1, 2, \dots, n-1). \quad (46)$$

Since $[\sigma^{ik}]$ must be Gramian if it exists, the last member of (45) shows α to be a quadratic form over this nonsingular matrix, so it must be that

$$\alpha > 0. \quad (47)$$

For $k = 1$, (46) shows that

$$\alpha_1 = \alpha b_1. \quad (48)$$

Differencing both members of (46) with respect to k shows further that

$$\alpha_k = \alpha \Delta_k b_{k-1} \quad (k = 2, 3, \dots, n-1); \quad (49)$$

and finally, for $k = n - 1$, (46) shows that $\alpha - \alpha_n = \alpha b_{n-1}$, or

$$\alpha_n = \alpha(1 - b_{n-1}). \quad (50)$$

Therefore, if we let

$$g_k = \begin{cases} b_1 & (k = 1) \\ \Delta_k b_{k-1} & (k = 2, 3, \dots, n-1) \\ 1 - b_{n-1} & (k = n) \end{cases} \quad (51)$$

we can write our desired formula compactly as

$$\alpha_k = \alpha g_k \quad (k = 1, 2, \dots, n). \quad (52)$$

It also follows from (51) that

$$\sum_{k=1}^n g_k = 1, \quad (53)$$

so (52) cannot be used to obtain an explicit formula for α .

An explicit formula for α is easily obtained as follows. Multiply both members of (38) by σ_{kl} and sum over l . Recalling (34), (44), and (52), we see that—changing subscripts—

$$\alpha \sum_{j=1}^n g_j \sigma_{jk} = 1 \quad (k = 1, 2, \dots, n), \quad (54)$$

or

$$\alpha = 1 / \left(\sum_{j=1}^n g_j \sigma_{jk} \right) \quad (k = 1, 2, \dots, n). \quad (55)$$

Using notation (51), and shifting notation from $k + 1$ to k , we can now rewrite (43) as

$$\sigma^{kl} = \alpha g_k g_l - \Delta_k (c_{k-1} \Delta_k \delta_{k-1,l}) \quad \begin{pmatrix} k = 2, 3, \dots, n-1 \\ l = 1, 2, \dots, n \end{pmatrix}. \quad (56)$$

Now, (56) gives all the elements of $[\sigma^{ik}]$ directly except for the first and last rows ($k = 1$ and $k = n$). These "boundary conditions" are obtained from (42). Setting $k = 1$ and using notation (50) show that

$$\sigma^{1l} = \alpha g_1 g_l - c_1 \Delta_1 \delta_{1,l} \quad (l = 1, 2, \dots, n). \quad (57)$$

Setting $k = n - 1$ in (42), and using (38) and (51), show that

$$\sigma^{nl} = \alpha g_n g_l + c_{n-1} \Delta_{n-1} \delta_{n-1,l} \quad (l = 1, 2, \dots, n). \quad (58)$$

IX. *The Inverse Matrix and the Ranks of Its Parts*

To see more graphically what the inverse matrix defined by (56), (57), and (58) looks like, let c_{kl} be defined, for all l , as

$$c_{kl} = \begin{cases} -c_1 \Delta_1 \delta_{1,l} & (k = 1) \\ -\Delta_k (c_{k-1} \Delta_k \delta_{k-1,l}) & (k = 2, 3, \dots, n-1) \\ c_{n-1} \Delta_{n-1} \delta_{n-1,l} & (k = n). \end{cases} \quad (59)$$

The right member of (59) expands into the following explicit statement of the elements of $[c_{kl}]$:

$$[c_{kl}] = \begin{bmatrix} c_1 & -c_1 & & & \\ -c_1 & c_1 + c_2 & -c_2 & & \\ & -c_2 & c_2 + c_3 & & \\ & & & \dots & \\ & & & & -c_{n-2} & c_{n-2} + c_{n-1} & -c_{n-1} \\ & & & & -c_{n-1} & c_{n-1} \end{bmatrix}. \quad (60)$$

$[\sigma^{kl}]$ can now be regarded as the sum of two matrices, for we can write

$$\sigma^{kl} = \alpha g_k g_l + c_{kl} \quad (k, l = 1, 2, \dots, n). \quad (61)$$

Now, from (59) and (44)—or from (60)—

$$\sum_{l=1}^n c_{kl} = 0 \quad (k = 1, 2, \dots, n), \quad (62)$$

or the columns (rows) of $[c_{kl}]$ are linearly dependent and the matrix is singular. It has been proved in (2) that the latent roots of a matrix of the form of (60) must all be distinct. So although $[c_{kl}]$ is singular, it can have only one zero root and hence must be of rank $n - 1$. By inspection of (60) it is seen that all submatrices with elements all on one side of, or on, the main diagonal are either of rank zero or one.

On the other hand, the matrix $[\alpha g_k g_l]$ is at most of rank one, being the product of a vector and its transpose. Indeed, for $[\sigma_{ik}]$ to be nonsingular, $[\alpha g_k g_l]$ cannot vanish, else the right member of (61) will be left only with the singular $[c_{kl}]$. So a necessary condition for the inverse to exist is that $[\alpha g_k g_l]$ be precisely of rank one, or that $[g_k] \neq 0$. This last condition is always assured by (53).

From the conclusions of the last two paragraphs, it is apparent that Theorem 2 holds for $[\sigma^{ik}]$ as well as for $[\sigma_{ik}]$. (Indeed, the writer has an unpublished theorem that shows a general correspondence between ranks of

parts of an inverse and the corresponding parts of the original matrix for any nonsingular matrix. We have merely worked out a special case here.)

X. Implications for Statistical Prediction

While the one-sided rank 2 condition holds for $[\sigma^{ik}]$, it is the details that are important. In general, the elements of an inverse matrix depend on all the elements of the original matrix, and will change as the order of the matrix is increased or decreased. But in (61), the *only over-all factor that changes as variables are added to or removed from the battery is α* , or the sum of all the elements in the inverse.

A coefficient g_k , as defined by (51) and (41), depends only on the variances and covariances of s_k and its immediate neighbors s_{k-1} and s_{k+1} . A coefficient c_{kl} will vanish, according to (60), unless $l = k - 1, k$, or $k + 1$. Therefore, if an s_{n+1} is added beyond the s_n of the given simplex, none of these coefficients will change except those for s_n . Or if a point is inserted between s_k and s_{k+1} in the simplex order, this will change coefficients associated only with points in the neighborhood of this new point. Thus again, as discussed in great detail in (3), in the multiple linear regression of any s_k on the remaining $n - 1$ distinct variables of the simplex, the multiple regression weights and multiple correlation coefficients depend essentially on the law of neighboring of the points of the simplex.

Again, the possibility appears that s_k can be essentially as predictable from s_{k-1} and s_{k+1} as it is from all the $n - 1$ distinct variables in the simplex apart from itself. We shall now see how, under certain circumstances, σ^{kl} is determined largely by c_{kl} and hardly by $\alpha g_k g_l$.

Specifically, we shall prove the following theorem:

THEOREM 3. *If λ_0 is the smallest latent root of $[\sigma_{ik}]$, then*

$$\alpha \leq 1 / \left(\lambda_0 \sum_{i=1}^n g_i^2 \right). \quad (63)$$

If $\lambda_0 \sum_{i=1}^n g_i^2 \rightarrow \infty$ as $n \rightarrow \infty$, then $\alpha \rightarrow 0$.

For the proof, multiply both members of (54) by g_k , sum over k and use (53) to see that

$$\alpha \sum_{i=1}^n \sum_{k=1}^n g_i g_k \sigma_{ik} = 1. \quad (64)$$

Now, the value λ_0 is the smallest obtainable by the quadratic form on the left of (64) when the g_i are normalized, or

$$\sum_{i=1}^n \sum_{k=1}^n g_i g_k \sigma_{ik} \geq \lambda_0 \sum_{i=1}^n g_i^2. \quad (65)$$

Hence, (63) follows from (64) and (65), and the theorem is established.

In circumstances where the g_i are approximately equal among themselves, quantities of the form $g_i g_k / \sum_{i=1}^n g_i^2$ will be of the order of $1/n$. Then, if the smallest root λ_0 does not tend to zero with n , or if it tends to zero at a slower rate than the order of $1/n$, it follows from Theorem 3 that $[\alpha g_i g_k] \rightarrow 0$ as $n \rightarrow \infty$, or from (61), $[\sigma^{kl}] \rightarrow [c_{kl}]$.

To have $[\alpha g_k g_l] \rightarrow 0$ would be a special case of an ϵ -simplex as defined in (3). The general definition of an ϵ -simplex is essentially non-parametric for finite n , in the sense that it is concerned only with limits as $n \rightarrow \infty$. It simply states that multiple regression coefficients should tend to zero for non-neighboring tests, or elements more than one diagonal away from the main diagonal of the inverse matrix should tend to zero as n increases. The simplex defined by law (21) can, therefore, be a special kind of ϵ -simplex.

XI. The Difference Equations for Principal Components

Having an explicit formula such as (61) for the inverse matrix helps us also to study the principal components defined by (31). Multiply both members of (31) by σ^{kl} and sum over k to obtain—revising subscripts—

$$a_k u_k = \lambda \sum_{i=1}^n u_i \sigma^{ik} \quad (k = 1, 2, \dots, n). \quad (66)$$

Let β be defined by

$$\beta = \sum_{k=1}^n g_k u_k. \quad (67)$$

Then using (61) and (67) in (66) shows that

$$a_k u_k = \lambda \left(\alpha \beta g_k + \sum_{i=1}^n u_i c_{ik} \right) \quad (k = 1, 2, \dots, n). \quad (68)$$

Now, the summation on the right is also expressible as first- and second-order differences among the c_k . For, using (59), we see that

$$\sum_{i=1}^n u_i c_{ik} = \begin{cases} -c_1 \Delta_1 u_1 & (k = 1) \\ -\Delta_l (c_{l-1} \Delta_l u_{l-1}) & (k = 2, 3, \dots, n-1) \\ c_{n-1} \Delta_{n-1} u_{n-1} & (k = n). \end{cases} \quad (69)$$

Thus, (68) can be regarded as expressing a second-order difference equation with two first-order boundary conditions.

Strictly speaking, however, more than a second-order difference equation is involved in (68), for β depends on *all* n of the u_i , according to (67). However, if $\beta = 0$, then (68) certainly reduces to the right order. If $\beta \neq 0$, we can divide both members of (68) through by β and regard the unknown to be u_k/β instead of u_k . Since the u_k are determined only up to a constant of

proportionality in any event, this is one way of taking up this degree of freedom.

The properties of the solutions to (68) in the general case remain to be explored. The previous special case of the restricted simplex in (3) is where $g_1 = 1$ and $g_j = 0$ for $j = 2, 3, \dots, n$. Then β in (67) is simply $\beta = g_1 u_1$ and (68) is

$$a_k u_k = \lambda \sum_{j=1}^n u_j (c_{jk} + \delta_{1j} \delta_{1k} \alpha g_1^2) \quad (k = 1, 2, \dots, n). \quad (70)$$

The matrix implied by the parentheses on the right differs from $[c_{jk}]$ in (60) merely by adding the quantity αg_1^2 to c_{11} , or the element c_1 in the first row and column. This merely changes the first boundary condition, as obtainable from (69), but leaves the rest of (69) unchanged.

The solutions to (70) have the law of oscillation discussed in (2) and (3).

Another special case of interest is where g_j is constant for all j . From (53), this implies

$$g_j = 1/n \quad (j = 1, 2, \dots, n). \quad (71)$$

If in addition, weights are chosen so that a_j is constant for all j , say

$$a_j = a \quad (j = 1, 2, \dots, n), \quad (72)$$

then it is easily seen from (68), (67) and (62) that $[g_j]$ is a latent vector with latent root $\lambda = na/\alpha$. Since all other latent vectors must be orthogonal to this one, it follows from (67) that $\beta = 0$ for the remaining latent vectors, and we are back to our standard type of difference equation for these remaining vectors; they are the vectors of $[c_{jk}]$. Hypotheses (71) and (72) lead to the case where the centroid is the same as a latent vector.

This raises the following question. If a resolution into components is desired, why not work in any case with those indicated by the formula for the inverse matrix? Certainly, basic structure properties are revealed by (61). If we again assume (72), then the first centroid loadings of (61) are $\sqrt{\alpha g_k}$ ($k = 1, 2, \dots, n$). According to the general formulas of (9) and (10), any Gramian matrix can have its rank reduced by extracting a centroid—the process is not restricted to correlation matrices and can be used on $[\sigma^{kl}]$ in particular. If we subtract out the contributions of these loadings from $[\sigma^{kl}]$, then we are left with the matrix of rank $n - 1$, $[c_{kl}]$, which now has an interesting law of principal components.

The factoring law suggested by this, then, is first to remove the first centroid, and then resolve the rest into principal components.

Since we are not factoring $[\sigma_{jk}]$ here but its inverse, we are not factoring the observed scores. Rather, by implication we are factoring the *anti-image* scores, for $[\sigma^{kl}]$ is closely related to the covariances among the anti-images of the s_j (6). That factoring a Gramian matrix is equivalent to factoring a

score matrix of which it is the product has been proved in (9) and discussed also in (10).

If (72) does not hold, then a more general weighted average is called for than the centroid to remove the term in $\alpha g_k g_l$ from (61).

XII. *The Sufficiency of the Formula for the Inverse Matrix.*

Up until now, we have not mentioned a somewhat important question. Under what conditions does (61) provide a matrix that is actually inverse to $[\sigma_{ik}]$? We have arrived at (61) by assuming $[\sigma_{ik}]$ to be nonsingular and (40) to hold. But can $[\sigma_{ik}]$ be nonsingular if law (21) holds? Fully to establish (61), we must prove that (34) actually holds assuming that $[\sigma_{ik}]$ obeys law (21), or (36).

A first indispensable assumption clearly is that (40) holds, for if two points coincide, $[\sigma_{ik}]$ must obviously be singular. Next, let us examine the assertion in (55) that the sums of the rows of $[\sigma_{ik}]$, when weighted by the g_i , are constant. Let h_k be defined as

$$h_k = \sum_{i=1}^n g_i \sigma_{ik} \quad (k = 1, 2, \dots, n). \quad (73)$$

Differencing both members of (73) with respect to k and using (36) yield

$$\begin{aligned} \Delta_k h_k &= (\sigma_{k,k+1} - \sigma_k^2) \sum_{i=1}^k g_i \\ &+ (\sigma_{k+1}^2 - \sigma_{k,k+1}) \sum_{i=k+1}^n g_i \quad (k = 1, 2, \dots, n-1). \end{aligned} \quad (74)$$

From (51) and (53),

$$\sum_{i=1}^k g_i = b_k, \quad \sum_{i=k+1}^n g_i = 1 - b_k \quad (k = 1, 2, \dots, n-1). \quad (75)$$

Multiply both members of (74) by c_k , use (75) and notation (41)—remembering (39)—to see that

$$c_k \Delta_k h_k = (b_k - 1)b_k + b_k(1 - b_k) = 0 \quad (k = 1, 2, \dots, n-1). \quad (76)$$

Therefore $\Delta_k h_k = 0$ for all possible k , or the h_k are constant.

Should the constant value of h_k be zero, then $[\sigma_{ik}]$ would be singular, according to the resulting linear dependence expressed by (73). Therefore, for the inverse to exist, we must assume the constant value of h_k to be different from zero. Let us denote this constant value by $1/\alpha$, or define α by (55).

We can now go ahead to define a matrix $[\sigma^{ki}]$ by (61), and proceed to

prove that it satisfies (34). Multiply both members of (61) by σ_{jk} , sum over k , and use (55) to see that

$$\sum_{k=1}^n \sigma_{jk} \sigma^{kl} = g_l + \sum_{k=1}^n \sigma_{jk} c_{kl} \quad (j, l = 1, 2, \dots, n). \quad (77)$$

In (59), since $c_{kl} = c_{lk}$, interchange k and l ; multiply through by σ_{jk} , and sum over k to obtain

$$\sum_{k=1}^n \sigma_{jk} c_{kl} = \begin{cases} -c_1 \Delta \sigma_{j1} & (l = 1) \\ -\Delta \frac{(c_{l-1} \Delta \sigma_{j, l-1})}{l} & (l = 2, 3, \dots, n-1) \\ c_{n-1} \Delta \frac{\sigma_{j, n-1}}{n-1} & (l = n). \end{cases} \quad (78)$$

Using (36) in (78), remembering notation (41) and (51), shows that

$$\sum_{k=1}^n \sigma_{jk} c_{kl} = \delta_{jl} - g_l \quad (j, l = 1, 2, \dots, n). \quad (79)$$

Substituting (79) into (77) shows that (34) holds, which is what was to be proved. These results can be summarized as a theorem.

THEOREM 4. *If $[\sigma_{jk}]$ satisfies law (21), then a necessary and sufficient condition for it to be nonsingular is that $d_{k, k-1} > 0$ ($k = 1, 2, \dots, n-1$) and that $\sum_{j=1}^n g_j \sigma_{jk}$ be different from zero for at least one value of k . Then $[\sigma_{jk}]^{-1}$ is given by formula (61).*

REFERENCES

1. Gabriel, R. K. The simplex structure of the Progressive Matrices test. *British J. of statist. Psychology*, 1954, 7, 9-14.
2. Guttman, L. The principal components of scale analysis. In Samuel A. Stouffer, et al., *Measurement and prediction*, Princeton Univ. Press, 1950.
3. Guttman, L. A new approach to factor analysis: the radex. In Lazarsfeld, Paul F. (editor), *Mathematical thinking in the social sciences*, The Free Press, 1954.
4. Guttman, L. Two new approaches to factor analysis. Annual technical report on project Nonr-731(00), submitted to the Office of Naval Research, Washington, D. C., by International Public Opinion Research, Inc. and the Israel Institute of Applied Social Research, June 1953.
5. Guttman, L. The principal components of scalable attitudes. In Lazarsfeld, Paul F. (editor), *Mathematical thinking in the social sciences*, The Free Press, 1954.
6. Guttman, L. Image theory for the structure of quantitative variates. *Psychometrika*, 1953, 18, 277-296.
7. Guttman, L. Some necessary conditions for common-factor analysis. *Psychometrika*, 1954, 19, 149-161.
8. Guttman, L. An additive metric from all the principal components of a perfect scale. *British J. statist. Psychol.*, 1955, 8, (In press).

9. Guttman, L. General theory and methods for matrix factoring. *Psychometrika*, 1944, 9, 1-16.
10. Guttman, L. Multiple group methods for common-factor analysis: their basis, computation, and interpretation. *Psychometrika*, 1952, 17, 209-222.
11. Thurstone, L. L. Primary mental abilities. *Psychometric Monographs No. 1*, Univ. Chicago Press, 1938.
12. Thurstone, L. L. and Thurstone, T. G. Factorial studies of intelligence. *Psychometric Monographs No. 2*, Univ. Chicago Press, 1941.

Manuscript received 6/1/54

Revised manuscript received 9/13/54

EQUATING TEST SCORES—A MAXIMUM LIKELIHOOD SOLUTION

FREDERIC M. LORD

EDUCATIONAL TESTING SERVICE

Certain problems of equating are discussed. The maximum likelihood solution is presented for the following special equating problem: Two tests, U and V , are to be equated, making use of a third "anchor" test, W . The examinees are divided into two random halves. Tests U and W are administered to one half; tests V and W are administered to the other half. It is assumed that any practice effect or other effect, exerted by U and V on W , is the same for U and for V .

Two tests may be said to be equated for a given group when the score scales on the two tests are so adjusted that both tests have the same frequency distribution of true scores in the given group. [Flanagan (1) and Gulliksen (2, pp. 296-304) give brief discussions of various methods of equating.] If the tests are equally reliable, then both tests will also have approximately the same frequency distribution of actual scores. As an approximation, two equally reliable tests may be equated by changing the score scale on either test in such a way that the distribution of actual scores becomes the same for both tests. The equipercentile method of equating is commonly used for this purpose.

If we wish to equate two equally reliable and otherwise approximately parallel forms of the same test, it is often convenient to assume that the score distributions of the two forms may differ somewhat in mean and variance, but that any other differences in the shape of these distributions may be ignored in practice. Under this assumption, the tests can be equated by simply changing the origin and the size of the unit of measurement of either score scale. If x and y are scores on two tests, the standardized scores $(x - \mu_x)/\sigma_x$ and $(y - \mu_y)/\sigma_y$ (μ and σ denote mean and standard deviation in the population of examinees for which the tests are to be equated) both have zero mean and unit variance; consequently, under the assumption outlined, standardized scores are equated, by definition.

Under the assumption of the foregoing paragraph, which will be implicit in all that follows, the only practical problem is to estimate μ_x , μ_y , σ_x , and σ_y for the population in which the two tests are to be used, so that the scores on both tests can be standardized. An obvious procedure is to administer test X to one random sample from this population and test Y to another random sample, and to estimate the desired parameters from the usual sample statistics.

The procedure just outlined, however, is not very efficient, since chance fluctuations produce differences in ability between the two groups, and these differences cause a bias in the equating. A more efficient method, provided the practice effect is properly handled, is to administer both tests to each examinee. Unfortunately, it is frequently not possible in practice to obtain sufficient testing time to administer two full-length tests to each examinee. A compromise procedure, suggested by Ledyard Tucker and commonly used at Educational Testing Service, is to divide the examinees into two random samples, each of which takes only one form of the test to be equated, and each of which also takes the same "anchor test." If this anchor test correlates highly with the other tests, its use greatly reduces the sampling errors of both the estimates obtained and the resulting equated scores. [Standard error formulas for equated scores obtained by the methods discussed here and by certain other methods are given in (4).]

It might be thought that the best procedure would be to equate each of the two forms to the anchor test and thus to each other. Actually, this procedure is inefficient, yielding estimates that, in certain cases, have considerably larger sampling errors than those obtained by ignoring the anchor test. An optimum equating procedure for handling the data in question is found by using the maximum likelihood method of estimation. The necessary estimates are derived, and the optimum procedure is outlined in what follows. The formulas that will be obtained differ only slightly from those used in Tucker's procedure, as discussed by Gulliksen (2, pp. 299-304); the assumptions made in reaching the present formulas are somewhat different.

Problem

Two tests, U and V , are to be equated, making use of a third anchor test, W . The examinees are divided into two random halves, which will be called the " a -group" and the " b -group." Tests U and W are administered to the a -group; tests V and W are administered to the b -group. It is assumed that any practice effect or other effect exerted by U and V on W is the same for U and for V .

R. S. Levine and W. Angoff (personal communication) have shown that the solution given here is also applicable when test W is a part of U and of V , i.e., tests U and V have common items W .

Notation

Consideration will be limited to the case where there are N examinees in each half-group. Let u_a and w_a denote the scores of examinee a , who is in the a -group, on tests U and W ; let v_b and w_b similarly denote scores of examinee b , who is in the b -group.

The symbols μ , σ , and ρ will be used to represent means, standard deviations, and correlation coefficients, respectively, in the population. The

population referred to here and in what follows is the population of all examinees from which the a - and b -group may be considered to be random samples.

Sample means will be denoted by \bar{u} , \bar{v} , \bar{w} ; sample standard deviations and correlations by s and r with appropriate subscripts. Where the meaning would otherwise be unclear, a single prime or a double prime will denote a statistic from the a -group or from the b -group, respectively.

Assumption

It is assumed that the scores u , v , and w have a normal trivariate distribution in the population. The joint distribution of u_a and w_a is thus

$$f_a(u_a, w_a) = \frac{1}{2\pi\sigma_u\sigma_w\sqrt{1-\rho_u^2}} \exp \left[-\frac{1}{2(1-\rho_u^2)} \left\{ \frac{(u_a - \mu_u)^2}{\sigma_u^2} + \frac{(w_a - \mu_w)^2}{\sigma_w^2} - 2\rho_u \frac{(u_a - \mu_u)(w_a - \mu_w)}{\sigma_u\sigma_w} \right\} \right], \quad (1)$$

ρ_u being the correlation between u and w . The joint distribution of v_b and w_b , denoted by $f_b(v_b, w_b)$, is the same as the foregoing except that u is replaced by v and a by b .

The Likelihood Function

The *likelihood* of occurrence of the actually observed values of u_a and w_a in the a -group is, by definition, $\prod_{a=1}^N f_a(u_a, w_a)$. Similarly, the *likelihood* for the b -group is $\prod_{b=1}^N f_b(v_b, w_b)$. The product of these two is the *likelihood function* (L) for all observed values in the data at hand. It will be convenient to work with the logarithm of the likelihood function, which is readily found to be

$$\begin{aligned} \log L = & -2N \log 2\pi - N \log \sigma_u\sigma_w \\ & - 2N \log \sigma_w - \frac{1}{2}N \log (1 - \rho_u^2)(1 - \rho_v^2) \\ & - \frac{1}{2(1 - \rho_u^2)} \left[\frac{1}{\sigma_u^2} \sum_a (u_a - \mu_u)^2 + \frac{1}{\sigma_w^2} \sum_a (w_a - \mu_w)^2 \right. \\ & \left. - \frac{2\rho_u}{\sigma_u\sigma_w} \sum_a (u_a - \mu_u)(w_a - \mu_w) \right] - \frac{1}{2(1 - \rho_v^2)} \left[\frac{1}{\sigma_v^2} \sum_b (v_b - \mu_v)^2 \right. \\ & \left. + \frac{1}{\sigma_w^2} \sum_b (w_b - \mu_w)^2 - \frac{2\rho_v}{\sigma_v\sigma_w} \sum_b (v_b - \mu_v)(w_b - \mu_w) \right]. \end{aligned} \quad (2)$$

The likelihood function contains eight unknown population parameters: $\mu_u, \mu_v, \mu_w, \sigma_u, \sigma_v, \sigma_w, \rho_u, \rho_v$. We wish to choose values of these parameters that will maximize the likelihood of occurrence of the actually observed sample. Consequently, we differentiate (2) with respect to each parameter

in turn and set each derivative equal to zero, at the same time placing a circumflex above the symbol for each parameter to indicate that we are now dealing with estimates of the parameters rather than with their true values. Eight simultaneous equations in eight unknowns are thus obtained.

The Likelihood Equations

After some cancellation and rearrangement, the first three equations are

$$\hat{\mu}_u - \hat{\beta}_{uw}\hat{\mu}_w = \bar{u} - \hat{\beta}_{uw}\bar{w}', \quad (3)$$

$$\hat{\mu}_v - \hat{\beta}_{vw}\hat{\mu}_w = \bar{v} - \hat{\beta}_{vw}\bar{w}'', \quad (4)$$

$$\frac{\hat{\mu}_w - \hat{\beta}_{uw}\hat{\mu}_u}{\hat{\kappa}_u^2} + \frac{\hat{\mu}_w - \hat{\beta}_{vw}\hat{\mu}_v}{\hat{\kappa}_v^2} = \frac{\bar{w}' - \hat{\beta}_{uw}\bar{u}}{\hat{\kappa}_u^2} + \frac{\bar{w}'' - \hat{\beta}_{vw}\bar{v}}{\hat{\kappa}_v^2}, \quad (5)$$

where $\hat{\kappa}_u^2 = 1 - \hat{\rho}_u^2$, $\hat{\kappa}_v^2 = 1 - \hat{\rho}_v^2$, and each $\hat{\beta}$ is a regression coefficient—for example, $\hat{\beta}_{uw} = \hat{\sigma}_u\hat{\rho}_u/\hat{\sigma}_w$.

Multiplying (3) by $\hat{\beta}_{uw}/\hat{\kappa}_u^2$, multiplying (4) by $\hat{\beta}_{vw}/\hat{\kappa}_v^2$, and adding both products to (5), we obtain, after simplification,

$$\hat{\mu}_w = \bar{w}, \quad (6)$$

where $\bar{w} = \frac{1}{2}(\bar{w}' + \bar{w}'')$ is the observed mean of w in the combined a - and b -group. Equation 6 presents the maximum likelihood estimate of μ_w .

Substituting (6) in (3) and in (4), we obtain, after simplification,

$$\hat{\mu}_u = \bar{u} - \hat{\beta}_u D, \quad (7)$$

$$\hat{\mu}_v = \bar{v} + \hat{\beta}_v D, \quad (8)$$

where $D = \frac{1}{2}(\bar{w}' - \bar{w}'')$, $\hat{\beta}_u$ is written for $\hat{\beta}_{uw}$, and $\hat{\beta}_v$ for $\hat{\beta}_{vw}$. These equations will be of practical use as soon as expressions have been found for $\hat{\beta}_u$ and $\hat{\beta}_v$.

The remaining five maximum likelihood equations are readily found to be

$$\hat{\sigma}_u^2 = (S_u^2 - \hat{\beta}_u C_{uw})/\hat{\kappa}_u^2, \quad (9)$$

a similar equation for v instead of u ,

$$2\hat{\sigma}_w^2 - \frac{1}{\hat{\kappa}_u^2} \left(S_w'^2 - \frac{1 - \hat{\kappa}_u^2}{\hat{\beta}_u} C_{uw} \right) - \frac{1}{\hat{\kappa}_v^2} \left(S_w''^2 - \frac{1 - \hat{\kappa}_v^2}{\hat{\beta}_v} C_{vw} \right) = 0, \quad (10)$$

$$\hat{\rho}_u - \frac{\hat{\rho}_u}{\hat{\kappa}_u^2} \left[\frac{S_w'^2}{\hat{\sigma}_w^2} + \frac{1}{\hat{\sigma}_u^2} (S_u^2 - 2\hat{\beta}_u C_{uw}) \right] + \frac{C_{uw}}{\hat{\sigma}_u \hat{\sigma}_w} = 0, \quad (11)$$

and a fifth equation like (11) but with v instead of u . In the foregoing equations,

$$S_u^2 = \sum_a (u_a - \hat{\mu}_u)^2/N, \quad (12)$$

$$C_{vw} = \sum_b (v_b - \hat{\mu}_v)(w_b - \hat{\mu}_w)/N, \quad (13)$$

and so forth.

Multiply (9) by $\hat{\rho}_u/\hat{\sigma}_u^2$ and subtract from (11) to obtain the result

$$\frac{\hat{\rho}_u}{\hat{\kappa}_u^2} \left(\frac{S'_{w^2}}{\hat{\sigma}_w^2} - \frac{\hat{\beta}_u C_{uw}}{\hat{\sigma}_u^2} \right) - \frac{C_{uw}}{\hat{\sigma}_u \hat{\sigma}_w} = 0. \quad (14)$$

Multiply (14) by $\hat{\kappa}_u^2$, write out $\hat{\kappa}_u^2$ and $\hat{\beta}_u$ in terms of $\hat{\rho}_u$, and simplify to obtain

$$S'_{w^2} \hat{\sigma}_u \hat{\rho}_u = C_{uw} \hat{\sigma}_w. \quad (15)$$

This may be rewritten

$$\hat{\beta}_u = C_{uw}/S'_{w^2}. \quad (16)$$

By a well-known formula, (12) can be rewritten

$$S_u^2 = s_u^2 + (\bar{u} - \hat{\mu}_u)^2, \quad (17)$$

where $s_u^2 = \sum_a (u_a - \bar{u})^2/N$ is the observed standard deviation of u . From (17) and (7),

$$S_u^2 = s_u^2 + \hat{\beta}_u^2 D^2. \quad (18)$$

Similarly,

$$S'_{w^2} = s'_{w^2} + D^2, \quad (19)$$

$$C_{uw} = c_{uw} + \hat{\beta}_u D^2, \quad (20)$$

and so forth, where $c_{uw} = \sum_a (u_a - \bar{u})(w_a - \bar{w})/N$ is the observed covariance of u and w in the a -group.

Substituting (19) and (20) into (16), we find after simplification

$$\hat{\beta}_u = c_{uw}/s'_{w^2}. \quad (21)$$

The expression on the right is the observed regression coefficient of u on w in the a -group, so we may write finally

$$\hat{\beta}_u = b_{uw}. \quad (22)$$

From (9), (18), (20), and (22)

$$\hat{\sigma}_{u \cdot w}^2 = (s_u^2 - b_{uw} c_{uw}) = s_u^2 (1 - r_{uw}^2) = s_{u \cdot w}^2, \quad (23)$$

where $\hat{\sigma}_{u \cdot w}^2 = \hat{\sigma}_u^2 \hat{\kappa}_u^2$, and $s_{u \cdot w}^2$ is the observed standard error of estimate in the a -group.

Finally, substitute (16) into (10) and simplify to obtain

$$\hat{\sigma}_w^2 = \frac{1}{2}(S'_{w^2} + S'_{w'^2}) = s_w^2, \quad (24)$$

where s_w^2 is the observed variance of w in the combined a - and b -group, i.e.,

$$s_w^2 = [(\sum_a w_a^2 + \sum_b w_b^2)/2N] - \bar{w}^2. \quad (25)$$

The writer is indebted to William H. Angoff for this simplified proof of (24).

The Maximum Likelihood Estimates

The set of eight equations, (3), (6), (22), (23), (24), and three equations in v analogous to those in u , is sufficient for the practical calculation of the maximum likelihood estimates of all the unknown parameters. A more convenient set of eight equations, readily derived from these, is

$$\hat{\mu}_w = \bar{w}, \quad (26)$$

$$\hat{\mu}_u = \bar{u}' + b'_{uw}(\bar{w} - \bar{w}'), \quad (27)$$

$$\hat{\mu}_v = \bar{v}'' + b''_{vw}(\bar{w} - \bar{w}''), \quad (28)$$

$$\hat{\sigma}_w^2 = s_w^2, \quad (29)$$

$$\hat{\sigma}_u^2 = \hat{\sigma}_{u \cdot w}^2 + \hat{\beta}_u^2 \hat{\sigma}_w^2 = s_u'^2 + b_{uw}'^2(s_w^2 - s_w'^2), \quad (30)$$

$$\hat{\sigma}_v^2 = s_v''^2 + b_{vw}''^2(s_w^2 - s_w''^2), \quad (31)$$

$$\hat{\sigma}_{uw} = \hat{\sigma}_u \hat{\sigma}_w \hat{\rho}_u = \hat{\beta}_u \hat{\sigma}_w^2 = b_{uw}' s_w^2, \quad (32)$$

$$\hat{\sigma}_{vw} = b_{vw}'' s_w^2, \quad (33)$$

where $\hat{\sigma}_{uw}$ and $\hat{\sigma}_{vw}$ are estimates of the population covariances. In the foregoing eight equations, primes or double-primes have been attached for the sake of clarity to all sample values except \bar{w} and s_w , these last two values being calculated from the combined a - and b -group.

(The maximum likelihood estimators presented in equations 26-33 constitute the solution of a general problem in estimating population parameters from incomplete data. A discussion of these results from this general point of view has been submitted for publication elsewhere.)

Equating

Granting the assumptions made from the start, a good equation for equating tests U and V is

$$(v - \mu_v)/\sigma_v = (u - \mu_u)/\sigma_u, \quad (34)$$

or, after rearranging,

$$v = Au + B, \quad (35)$$

where

$$A = \sigma_v/\sigma_u, \quad (36)$$

$$B = \mu_v - A\mu_u. \quad (37)$$

In (36) and (37), A and B are expressed in terms of the population parameters, which are unknown. We wish to use maximum likelihood estimates of A and B in (35). Since the maximum likelihood estimate of a certain

function of the parameters is the same as that function of the maximum likelihood estimates of the parameters, the equation to use for equating is

$$v = \hat{A}u + \hat{B}, \quad (38)$$

where $\hat{A} = \hat{\sigma}_v / \hat{\sigma}_u$, $\hat{B} = \hat{\mu}_v - \hat{A}\hat{\mu}_u$, the values of $\hat{\mu}_u$, $\hat{\mu}_v$, $\hat{\sigma}_u$, and $\hat{\sigma}_v$ being computed from the data by means of equations 27, 28, 30, and 31.

The formulas for equating thus obtained by the maximum likelihood method differ from those of Tucker, as discussed by Gulliksen, chiefly as a result of the fact that Tucker's procedure calls for estimating the performance of the *b*-group on test *U*, whereas the present procedure calls for estimating the performance of the entire population on both tests *U* and *V*. The present development is based on the assumption that the two groups tested are random samples from the same population. The assumptions made in Tucker's development do not require this, but they do impose considerable restriction on the nature of the differences between the two groups.

Numerical Example

The following illustrative example is based on real data taken from Karon's empirical study of equating methods (3). The raw data are given in the top half of Table 1; the necessary maximum likelihood estimates,

TABLE 1
Raw Data and Maximum Likelihood Estimates Needed for Equating

	Group <i>a</i>		Group <i>b</i>		Combined groups
	Test <i>U</i>	Test <i>W</i>	Test <i>V</i>	Test <i>W</i>	Test <i>W</i>
Mean (\bar{u} , \bar{v} , \bar{w})	117.85	34.36	115.33	33.42	33.89
Variance (s^2)	1129.62	116.81	1109.65	114.89	116.07
Regression on <i>w</i>	2.6744		2.6479		
$\hat{\mu}$	116.59		116.58		
$\hat{\sigma}^2$	1124.34		1117.92		

computed by equations (27), (28), (30), and (31), are given in the bottom half. Each group contains a random sample of 600 examinees. The final equation, obtained from (38),

$$v = .997u + 0.32, \quad (39)$$

gives the raw score (*v*) on test *V* that is equivalent to any given score (*u*) on test *U*.

If test W had not been administered, the final equation would have been

$$\begin{aligned} v &= (s_v/s_u)u + \bar{v} - (s_v/s_u)\bar{u} \\ &= .991u - 1.47. \end{aligned} \quad (40)$$

The use of test W provides the information that the b -group is probably slightly less competent and slightly less variable than the a -group (these differences having arisen solely because of sampling fluctuations). The maximum likelihood estimates in Table 1 and the resulting equation 39 take this sampling fluctuation into account, whereas equation 40 does not.

REFERENCES

1. Flanagan, J. C. Units, scores and norms. In E. F. Lindquist (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1950, pp. 695-763.
2. Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.
3. Karon, B. P. The stability of equated test scores. Research Bulletin 54-25. Princeton, N. J.: Educational Testing Service, 1954. (Multilithed)
4. Lord, F. M. Notes on comparable scales for test scores. Research Bulletin 50-48. Princeton, N. J.: Educational Testing Service, 1950. (Multilithed)

Manuscript received 5/12/54

Revised manuscript received 8/26/54

AXIOMS OF A THEORY OF DISCRIMINATION LEARNING*

FRANK RESTLE
STANFORD UNIVERSITY

Analysis of an empirical theory into a formal system with specified primitive notions and axioms has the advantage of making it clear what deductions from the theory are permissible, and clarifying the internal structure of the theory. An example of such analysis is presented in this paper.

Learning theories recently published by Estes and his associates (3, 4, 5) and Bush and Mosteller (1) have been characterized by mathematical formulation and reasoning. The writer has offered a similar theory designed for the analysis of two-choice discrimination learning. This new theory, using a strong simplifying assumption, yields several empirical predictions which have, in the main, been verified (9).

According to this theory the subject is faced on each trial with a collection of cues; some are relevant to getting reward and others irrelevant. On each trial of training some relevant cues are newly conditioned to the correct response and some irrelevant cues are newly adapted. A conditioned cue contributes to a correct response. An adapted cue becomes non-functional and does not directly affect the choice reaction.

The probability that a relevant cue will be conditioned on any trial (given that it has not been conditioned on a previous trial) will be denoted by θ . Since θ is constant from trial to trial and the same for all cues, the learning functions here are the same as the conditioning functions in the work of Estes and his associates (3, 5, and the "equal- θ approximation" case in 4.)

The fundamental assumption of the theory deals with θ . This assumption is that θ is the relative weight of relevant cues in the problem. The more relevant cues there are in the problem, the greater is the probability that any given relevant cue will be conditioned and that any given irrelevant cue will be adapted. By this simplifying assumption it is possible to make the theory unusually determinate.

In the earlier paper on this theory (9) a number of quantitative empirical

*This paper is adapted from part of a Ph.D. dissertation submitted to the Department of Psychology, Stanford University, in November 1953. The author wishes to express his appreciation to Dr. Patrick Suppes, who guided the analysis reported in this paper. The author is now with the Human Resources Research Office, The George Washington University.

laws were developed which were tested against experimental data. In general, the proposed laws were verified.

In the present paper a more precise and complete statement of the theory is made. Using only terms definable within the language of set-theory and logic, a complete list of primitive notions is given and the axioms are stated. Deductions are carried out entirely by the methods of formal mathematics, without recourse to psychological intuition or "good sense."

Before presenting the system it may be useful to describe the mathematical notions to be used. A *binary relation* is a relation between two entities. By a *set* is meant any arbitrary collection of things. In the formula $f(x, y) = z$, the term z is the value of the binary function f . If z is a real number, we say that f is *real-valued*. An *ordered couple* is a set which has two members, with the restriction that specifying the set requires not only naming the members but also indicating what order they come in. If $\langle x, y \rangle$ is an ordered couple and $x \neq y$, then $\langle x, y \rangle \neq \langle y, x \rangle$.

The usual set-theory notation is used; if X and Y are sets, $X \cup Y$ includes everything which is in either X or Y , $X \cap Y$ includes the elements which are in both X and Y , and $X - Y$ includes all the elements which are in X and are not in Y . The empty or null set is called Λ . In the body of the paper, capital letters are used to denote the sets and the one relation used; lower-case letters designate functions, integers, and variables. One function is given the designation θ to follow earlier usage (4, 5).

Primitive Notions

This system of discrimination learning is based on seven primitive notions, K , S^* , Q , w , c , a , and p . K is a set, S^* is a set of ordered couples, Q is a binary relation, w is a unary real-valued function, and c , a , and p are binary real-valued functions.

The set K is intended to be interpreted as the collection of cues. A cue is anything, concrete or abstract, present, past or future, of any description, to which the subject can learn to make a differential response. Obviously, at any given time there are cues to which the subject does not make responses—otherwise, there would be no learning. But if the subject can learn a differential response to something, by some training method, then that thing is a cue. Some cues are relatively simple energy sources. Some subjects can learn to respond to spatial or temporal patterns of objects or events; some produce reactions, overt, perceptual, or "thinking," which they can discriminate. Accounts of mediating processes can be found in work by Lawrence (6) and Wyckoff (10).

The set S^* is intended to be interpreted as any collection of two-choice discrimination problems, all of which involve the same pair of choice reactions. A problem S is uniquely associated with a pair of sets of cues: the set of

relevant cues, R , which the subject can use to predict reward, and the set of irrelevant cues, I , which are uncorrelated with reward and therefore cannot be used to predict reward.

If S is a problem in S^* and n is a positive integer, then SQn is interpreted as the statement, "problem S appears on trial n ." This is true if the subject must make a choice reaction in problem S on the n th trial.

If k is a cue, $w(k)$ is interpreted as the weight of cue k . According to Axiom D2, w is a discrete probability distribution defined over the class K of cues.

If k is a cue and n is a positive integer, then $c(k, n)$ is the probability that k is conditioned to the correct response at the beginning of the n th trial. If k is a cue then $a(k, n)$ is the probability that k is adapted at the beginning of the n th trial.

Before stating the axioms of this system we define $\theta(S)$ as the relative weight of relevant cues in problem S . This term will appear later in the learning functions of Axiom D7.

Definition: If $S = \langle R, I \rangle$ is in S^* , then $\theta(S) = \sum_{k \in R} w(k) / \sum_{k \in (R \cup I)} w(k)$.

Axioms

Definition: A system $\langle K, S^*, Q, w, c, a, p \rangle$ satisfying Axioms D1—D8 is called a *system of simple discrimination learning*.

AXIOM D1. K and S^* are non-empty, at most denumerable sets.

AXIOM D2. If k is in K , $w(k) \geq 0$, and $\sum_{k \in K} w(k) = 1$.

AXIOM D3. If $S = \langle R, I \rangle$ is in S^* , then R and I are subsets of K .

AXIOM D4. If $S = \langle R, I \rangle$ is in S^* , then the intersection of R and I is empty.

AXIOM D5. If S_1 and S_2 are distinct members of S^* , if n is a positive integer, and if S_1Qn , then not S_2Qn .

AXIOM D6. If $S = \langle R, I \rangle$ is in S^* , then for all k in $R \cup I$, $c(k, 1) = a(k, 1) = 0$.

AXIOM D7. If $S = \langle R, I \rangle$ is in S^* and n is a positive integer and SQn , then:

If k is in R , then $c(k, n+1) = c(k, n) + \theta(S)[1 - c(k, n)]$ and $a(k, n+1) = a(k, n)$.

If k is in I , $c(k, n+1) = c(k, n)$ and $a(k, n+1) = a(k, n) + \theta(S)[1 - a(k, n)]$.

Otherwise, $c(k, n+1) = c(k, n)$ and $a(k, n+1) = a(k, n)$.

AXIOM D8. If $S = \langle R, I \rangle$ is in S^* and n is a positive integer, then

$$p(S, n) = \frac{1}{2} \cdot \frac{\sum_{k \in (R \cup I)} w(k) - \sum_{k \in I} a(k, n) \cdot w(k) + \sum_{k \in R} c(k, n) \cdot w(k)}{\sum_{k \in (R \cup I)} w(k) - \sum_{k \in I} a(k, n) \cdot w(k)}.$$

Axiom D1 eliminates the trivial case in which either there are no cues or there is no problem and avoids mathematical difficulties by keeping K and S^* denumerable at most. Axiom D2 states that w is a discrete probability function. Axiom D3 states that the relevant and irrelevant cues in any problem are cues in the class K . Axiom D4 states that no cue can be both relevant and irrelevant in the same problem. Axiom D5 states that only one problem may occur on a given trial. Axiom D6 states that the system deals with a theoretically "naive" subject who, at the beginning of training

(trial 1), had neither conditioned nor adapted to any of the cues involved. Axiom *D7* states the laws of conditioning and adaptation, which are discussed above and in the earlier paper on this subject (9). Axiom *D8* states the "law of performance," giving p , the probability of a correct response, as a function of the number of conditioned and adapted cues. Inspection will show that p is the proportion of non-adapted (i.e., still-functional) cues which are conditioned plus one-half the proportion of non-adapted cues which are unconditioned.

Theorems

The theorems to be proved could not be proved rigorously with the system in (9). The equations derived in Theorems 2, 3, and 4 were compared directly with experiments.

The first empirical problem of the theory is the evaluation of the learning constant, $\theta(S)$, from discrimination learning data. This is accomplished by Theorem 1, which gives an explicit function relating $p(S, n)$ to $\theta(S)$. It is found in Corollary 1.1 that $p(S, n)$ is monotonic with respect to both $\theta(S)$ and n . Therefore, graphs can be constructed to determine θ knowing the empirical learning function, which corresponds to $p(S, n)$. Since such curve-fitting is unsatisfactory when dealing with individual subjects, and is invalid when dealing with groups of subjects who have different learning constants, we derive an explicit function relating the total number of errors expected, $\sum_{n=1}^{\infty} [1 - p(S, n)]$, to $\theta(S)$. Thus, if the learning experiment is continued until the subject has achieved a high criterion, the total errors made can be used to determine $\theta(S)$. Theorem 1 and its corollaries make it possible to evaluate $\theta(S)$ in practice.

The second empirical problem has to do with the combination of cues. Experimentally, we observe representative subjects learning to discriminate between, say, black and white, and from this we determine $\theta(S_{B-W})$. In the same apparatus we observe a second group of subjects learning to discriminate, for example, high and low pitches, and we determine $\theta(S_{H-L})$. The two sets of cues, brightness and pitch, are selected as ones which will not probably affect one another perceptually. In the same apparatus a third problem is run in which both brightness and pitch cues are relevant; for example, the subject must discriminate black and high pitch from white and low pitch. Theorem 2 makes it possible to predict $\theta(S_{B \text{ and } H-W \text{ and } L})$, and thus predict performance on this combined-cues problem (2).

The third empirical problem has to do with transfer of training from an easy discrimination problem to a more difficult one of the same sort. For example, a subject may be trained to approach *black* and avoid *white*, and is then trained to approach *dark gray* and avoid *light gray*. The experiment is interpreted as follows: we assume that the two problems present the same cues; the difference is that some of the cues which are relevant in the easier

problem are irrelevant in the more difficult problem. The more difficult problem is constructed from the easier one by shifting some cues from the set of relevant cues into the set of irrelevant cues. To predict transfer performance we first determine the θ values of the two problems by running them separately with naive subjects. From these values and knowledge of the number of trials of training on the easy problem, we can predict $p(S_{\text{hard}}, n)$ for all trials on the hard problem (7, 9). The required formula is derived in Theorem 3, and the total number of errors made in transfer is derived in Corollary 3.1.

The "converse" of the experiment discussed in Theorem 3 is an experiment in which the subject is first trained on the difficult problem and is then transferred to an easier one of the same sort (9). The formula for predicting performance, based on knowledge of $\theta(\text{easy})$, $\theta(\text{hard})$, and the number of pretraining trials on the hard problem, is given in Theorem 4.

Theorems 2, 3, and 4 make exact quantitative predictions of expected performance curves. Testing the predictions against empirical results does not involve curve-fitting and the use of arbitrary empirical constants. The predicted curve can in principle be drawn before any subjects are run on the test problem, and the theory is not confirmed unless the test performance corresponds to a particular learning curve predicted.

Since the proofs of the theorems are elementary in principle and somewhat tedious, only the method of proof will be given. The careful reader can verify for himself that entirely formal proofs are possible.

THEOREM 1. *If S is in S^* and SQj for all positive integers $j \leq n$, then [using θ as an abbreviation for $\theta(S)$]*

$$p(S, n) = 1 - \frac{1}{2}[(1 - \theta)^{n-1}]/[\theta + (1 - \theta)^n].$$

PROOF. We note that if k is in R , $c(k, n) = 1 - (1 - \theta)^{n-1}$ and if k is in I , $a(k, n) = 1 - (1 - \theta)^{n-1}$. The theorem is obtained by elementary algebra: the above values are substituted into Axiom D8, all terms are divided by $\sum_{k \in (R \cup I)} w(k)$, and the definition of θ is employed to simplify.

COROLLARY 1.1. *Under the conditions of Theorem 1, $p(S, n)$ is a monotonic non-decreasing function of n and a monotonic increasing function of θ .*

PROOF. This follows immediately from the theorem.

COROLLARY 1.2. *Under the above conditions,*

$$\sum_{n=1}^{\infty} [1 - p(S, n)] \cong \frac{1}{2} + \frac{1}{2}[\log \theta]/[(1 - \theta) \log (1 - \theta)].$$

PROOF. We first estimate $p(S, n)$ by the continuous function $p'(S, t) = 1 - \frac{1}{2}[(1 - \theta)^{t-1}]/[\theta + (1 - \theta)^t]$, and integrate $1 - p'(S, t)$ by using the substitution, $y = (1 - \theta)^t$.

THEOREM 2. If $S_1 = \langle R_1, I_1 \rangle$, $S_2 = \langle R_2, I_2 \rangle$ and $S_3 = \langle R_3, I_3 \rangle$ are in S^* and if $\sum_{k \in I_1} w(k) = \sum_{k \in I_2} w(k) = \sum_{k \in I_3} w(k)$, and if $\sum_{k \in R_1} w(k) + \sum_{k \in R_2} w(k) = \sum_{k \in R_3} w(k)$, then

$$1 - \theta(S_3) = [1 - \theta(S_1)][1 - \theta(S_2)]/[1 - \theta(S_1)\theta(S_2)].$$

PROOF. The proof follows immediately from the definition of θ .

THEOREM 3. (i) If $S_1 = \langle R_1, I_1 \rangle$ and $S_2 = \langle R_2, I_2 \rangle$ are in S^* and if R_2 is a subset of R_1 and I_1 is a subset of I_2 , and if

$$\sum_{k \in (R_1 \cup I_1)} w(k) = \sum_{k \in (R_2 \cup I_2)} w(k), \quad \text{and}$$

(ii) if for all $i \leq n$, $S_1 Q_i$, and for all $n + j$, $S_2 Q(n + j)$, then

$$p(S_2, n + j) = \frac{\theta_2 + \frac{1}{2}(1 - \theta_2)^{i-1}[\theta_1 - \theta_2 + (1 - \theta_1)^{n+1} - \theta_2(1 - \theta_1)^n]}{\theta_2 + (1 - \theta_2)^{i-1}[\theta_1 - \theta_2 + (1 - \theta_1)^{n+1}]}.$$

PROOF. Let k be a cue in R_2 . Since R_2 is a subset of R_1 , k is also in R_1 . At the beginning of trial $n + 1$, for all k in R_1 , $c(k, n + 1) = 1 - (1 - \theta_1)^n$. After $j - 1$ further trials on the second problem, $c(k, n + j) = 1 - (1 - \theta_1)^n (1 - \theta_2)^{j-1}$. This is the conditioning of all cues relevant in the second problem. At the beginning of trial $n + 1$, for all cues in I_1 , $a(k, n + 1) = 1 - (1 - \theta)^n$. For cues which are in I_2 but are not in I_1 , $a(k, n + 1) = 1 - (1 - \theta)^0 = 0$. (The fact that these latter cues have been conditioned is of no importance, since they are not relevant.) The theorem is obtained by using Axiom D7 to determine $c(k, n + j)$ and $a(k, n + j)$, substituting these values into Axiom D8, dividing by $\sum_{k \in (R \cup I)} w(k)$, and collecting like terms.

COROLLARY 3.1. Under the conditions of Theorem 3,

$$\sum_{j=1}^{\infty} [1 - p(S_2, n + j)] \cong \frac{B - A}{\theta_2 + B} + \frac{B - A}{B \log(1 - \theta_2)} [\log \theta_2 - \log(\theta_2 + B)],$$

where $A = \frac{1}{2}[\theta_1 - \theta_2 + (1 - \theta_1)^{n+1} - \theta_2(1 - \theta_1)^n]$, and $B = \theta_1 - \theta_2 + (1 - \theta_1)^{n+1}$.

PROOF. Note that by Theorem 3,

$$1 - p(S, n + j) = [\theta_2 + (B - A)(1 - \theta_2)^{j-1}]/[\theta_2 + B(1 - \theta_2)^{j-1}].$$

This is approximated by a continuous function, substituting the real variable t for j , and the resulting function is integrated, giving the corollary.

THEOREM 4. Given the same conditions as under (i) in Theorem 3, but if for all $i \leq n$, $S_2 Q_i$ and if for all $n + j$, $S_1 Q(n + j)$, then

$$p(S_1, n + j) = \frac{\theta_1 - \frac{1}{2}(1 - \theta_1)^{i-1}[\theta_1 - \theta_2 + \theta_2(1 - \theta_2)^n - (1 - \theta_2)^n(1 - \theta_1)]}{\theta_1 + (1 - \theta_2)^n(1 - \theta_1)^i}.$$

PROOF. The proof is similar to that of Theorem 3.

Discussion

Certain characteristics of the axiom system offered in this paper may require explanation. The extremely abstract nature of the axioms is designed to separate carefully the formal system from its psychological interpretation. This separation makes it possible to be sure that all needed assumptions have been explicitly stated. Axioms *D1*—*D6* are formal in nature and do not represent crucial psychological assumptions. However, if the required theorems are to be proved rigorously, such axioms are necessary (8).

The purpose of the paper is to make clear the formal assumptions, not their empirical consequences. However, it may be noted that if the four primitive notions *K*, *S**, *Q*, and *w* are defined operationally, the other three, *c*, *a*, and *p*, can be defined explicitly by using Axioms *D7* and *D8* as definitions. If the notion of a cue can be made clear, there is not likely to be any difficulty with the notion of a class of discrimination problems, or the occurrence of a problem. Operational definition of *w*, the weight or probability of a cue, seems at first glance difficult, but since the theory makes it possible to evaluate θ for any problem, one can in principle measure the ratio of weights of any two sets of cues. Thus, the measurement of *w* does not offer a theoretical difficulty, however complex the experimental manipulations may become.

The empirical definition of a cue is roughly the following: *k* is a cue if and only if, when the subject is given appropriate training, then he learns to make differential responses based solely on *k*. Here appropriate training is the most efficient training program possible. Often we do not know what training program this is or how long training must be continued to get learning, with the result that empirical use of this definition is hindered. It does, however, give a fairly clear intuitive idea of the meaning of the term *cue*.

To define *S**, the set of problems, it is essential only to know what a cue is and to distinguish relevant from irrelevant cues. A cue is relevant in a particular problem if it can be used in that problem as the basis for consistently correct response. A cue is irrelevant if the problem is so designed that the cue cannot be used as the basis for consistently correct response.

The relation of occurrence, *Q*, of a problem, does not take into account whether the subject makes a correct or incorrect response. Given the concept of a problem, the notion of occurrence of a problem is clear since it corresponds to the usual experimental notion of a trial (especially in non-correction type training where one run through the apparatus or situation is considered a trial).

Another characteristic of this theory is the very strong assumption identifying θ with the relative weight of relevant cues. Without this assumption it would have been extremely difficult to evaluate the needed learning parameters, and experimental tests would have been complicated immeasurably. While one may be skeptical that such a convenient assumption would

be satisfied, it permits a coherent and powerful theory to be constructed. Having made a very useful simplifying assumption, the theorist can always retreat when the data demand it.

Finally, it may be noted that this theory in its present form does not account for that important class of experiments in which the relevant cues are reversed, i.e., where the formerly correct cue becomes incorrect, and the formerly incorrect cue becomes correct. Generalization to this field of data is needed to broaden the empirical base of the theory.

REFERENCES

1. Bush, R. R. and Mosteller, F. A model for stimulus generalization and discrimination. *Psychol. Rev.*, 1951, **58**, 413-423.
2. Eninger, M. U. Habit summation in a selective learning problem. *J. comp. and physiol. Psychol.*, 1952, **45**, 604-608.
3. Estes, W. K. Toward a statistical theory of learning. *Psychol. Rev.*, 1950, **57**, 94-107.
4. Estes, W. K. and Burke, C. J. A theory of stimulus variability in learning. *Psychol. Rev.*, 1953, **60**, 276-286.
5. Estes, W. K. and Straughan, J. H. Analysis of a verbal conditioning situation in terms of statistical learning theory. *J. exp. Psychol.*, 1954, **47**, 225-234.
6. Lawrence, D. H. Acquired distinctiveness of cues: II. Selective association in a constant stimulus situation. *J. exp. Psychol.*, 1950, **40**, 175-188.
7. Lawrence, D. H. The transfer of a discrimination along a continuum. *J. comp. and physiol. Psychol.*, 1953, **45**, 511-516.
8. McKinsey, J. C. C., Sugar, A. C., and Suppes, P. Axiomatic foundations of classical particle mechanics. *J. rational Mechanics and Analysis*, 1953, **2**, 253-272.
9. Restle, F. A theory of discrimination learning. *Psychol. Rev.*, 1955, **62**, 11-19.
10. Wyckoff, L. B. The role of observing responses in discrimination learning. *Psychol. Rev.*, 1952, **59**, 431-442.

Manuscript received 6/9/54

Revised manuscript received 10/30/54

THE OBJECTIVE DEFINITION OF SIMPLE STRUCTURE IN LINEAR FACTOR ANALYSIS*

LEDYARD R. TUCKER
PRINCETON UNIVERSITY
AND
EDUCATIONAL TESTING SERVICE

Requirements for an objective definition of simple structure are investigated and a number of proposed objective criteria are evaluated. A distinction is drawn between exploratory factorial studies and confirmatory factorial studies, with the conclusion drawn that objective definition of simple structure depends on study design as well as on objective criteria. A proposed definition of simple structure is described in terms of linear constellations. This definition lacks only a statistical test to compare with possible chance results. A computational procedure is also described for searching for linear constellations. This procedure is very laborious and might best be accomplished on high-speed automatic computers. There is no guarantee that the procedure will find all linear constellations, but it probably would yield satisfactory results for well-designed studies.

The principle of simple structure, proposed by Thurstone as a solution to the problem of indeterminacy of position of axes in the factorial structure, has received wide support and use in factor analysis. There have been, however, a variety of criticisms including (1) a skepticism regarding whether this principle of simplicity did, in reality, adequately parallel nature, and (2) a feeling of disturbance at the subjectivity involved both in theory and in application. The first problem, that of the validity of the simple structure concept, may be settled only by experimental studies. It is the purpose of this paper to assist in solving the second problem, that of subjectivity, by attempting to develop a more objective and operational view of the simple structure concept.

Two major concepts of the nature of factors are used to justify the principle of simple structure. Thurstone's views might best be summarized by the following quotations: "In the interpretation of mind we assume that mental phenomena can be identified in terms of distinguishable functions, which do not all participate equally in everything that mind does . . . No

*This research was jointly supported by Princeton University, the Office of Naval Research under contract N6onr-270-20, and the National Science Foundation under grant NSF G-642. The author is especially indebted to Harold Gulliksen for his many exceedingly helpful comments and suggestions made during the course of this development. A debt of gratitude is also owed to Mrs. Gertrude Diederich, who performed many intricate calculations in the experiments on computing procedures. The author further wishes to express his appreciation to Frederic M. Lord and David R. Saunders, who read the manuscript and made a number of very useful suggestions.

assumption is made about the nature of these functions, whether they are native or acquired or whether they have a cortical locus." (14, p. 57.) "Just as we take for granted that the individual differences in visual acuity are not involved in pitch discrimination, so we assume that in intellectual tasks some mental or cortical functions are not involved in every task. This is the principle of 'simple structure' or 'simple configuration' in the underlying order for any given set of attributes." (14, p. 58.) Cattell (3) expresses a similar view. In contrast to the foregoing, Holzinger and Harman (5) express a variant view that factor analysis, as a branch of statistical analysis, conveys information in the original data with an aim of parsimony which should not be construed as a search for fundamental categories. Similarly, Vernon (19) takes a position that "... it should be clear that a factor is a construct which accounts for the objectively determined correlations between tests, in contrast to a faculty which is a hypothetical mental power." (19, p. 8.) Others have taken views on either of these two sides, with still others sticking to some middle ground. Since each of these views can be interpreted as yielding support for the desirability of simple structure, we believe that the definitions to follow could be derived from either view and will not distinguish between them. Some such view is necessary, however, as an initial step toward acceptance of the simple structure concept.

Relation Between Design of Factor Analysis Studies and Simple Structure

The factorial study of human behavior might best be conceived as a program of studies rather than in terms of isolated, separate studies. Each study should build upon the knowledge gained from previous studies and add further to the verified fund of knowledge. Early studies in some domain, or class of behavior, will be more exploratory in nature and be made with less perfected batteries of measures. As knowledge increases concerning the interrelations of the various behaviors in such a domain, it should be possible to construct more satisfactory batteries for factorial analysis. Confirmatory studies should aid in firmly establishing the factorial structure.

In exploratory studies a fully determined simple structure solution should not be expected and rotation of axes will probably be continued on subjective bases. There may well be an attempt to maximize the number of small, insignificant factor loadings; but some attention may also be given to interpretive possibilities. While some assistance may be obtained from analytic procedures, it seems inevitable that the rotation of axes for exploratory studies will remain an art. This paper does not attempt to present a method for rotation of axes to simple structure in exploratory studies. Rather, in contrast, the definitions and procedure to follow are to be conceived as applying primarily to the more perfected factorial studies.

A major premise of the present argument is that the objective definition

of a simple structure is dependent on both an adequate study design and on objective analytic criteria. Not all factorial studies may possess a simple structure, only those studies involving an appropriate battery of measures made on an appropriate sample of individuals. Some requirements set forth by the analytic criteria may be met only in the study design. It is desirable, however, that there be a maximum of freedom in the design of factorial studies so as to fit as many situations as possible. For example, an experimenter should be in a position to test objectively hypotheses concerning the relations of complex measures to factorially simpler ones. Thus, it is desirable that the analytic criteria permit complex variables and not limit the study design to factorially pure measures. The factorial simple structure needs to be unambiguously present, however, in the data. This is a function of the study design.

Requirements for Objective Definition of Simple Structure

Following is a proposed list of requirements for satisfactory objective criteria for simple structure. These requirements should be interpreted as applying to individual studies since invariance of factorial results over various changes in the population of individuals sampled and in the battery of measures is a matter for experimental verification. It will be noted, however, that small variations of factor loadings and projections from ideal values are permitted. These small variations from ideal might result either from random sampling error peculiar to the sample of individuals measured or from errors of approximation in the basic factorial model.

A second point to be noted is that a choice is made as to kind of projection employed relating test vectors to factors. In the case of correlated factors, orthogonal projections of test vectors on normals to hyperplanes are used. These orthogonal projections for a particular factor depend upon location of only the hyperplane for that factor and upon the test vectors. They are independent of the locations of all other hyperplanes. A further reason for this choice as to type of projection is that the square of this type of projection can be interpreted to represent the independent contribution of the factor to the variance of the variable.

a. Basic requirements

1. Emphasis is placed on a maximum concentration of vectors along hyperplanes, that is, on a maximum number of zero projections on normals to the hyperplanes, allowance being made for small variations in observed projections.
2. The vectors interpreted as being in each hyperplane span a space of $(r - 1)$ dimensions, allowance being made for small variations in observed projections, where r is the number of dimensions in the common-factor space.
3. Exactly as many simple structure factors are obtained as there are dimensions in the common-factor space.

b. *Types of freedom explicitly permitted*

4. Oblique factors are permitted.
5. A minority of highly complex measures whose vectors have projections on several, up to all, factors is permitted in the battery being analyzed.

c. *Operational requirements*

6. The choice as to which projections are to be interpreted as zero is made on objective grounds.
7. An objectively determined best fit to the data is involved.
8. The best fit is unbiased in the limiting sense that when the variance of projections interpreted as zero is small, the mean of these projections is near zero.
9. Statistical tests exist which indicate the plausibility of accepting any particular solution as a simple structure.
10. An automatic computational procedure is available for use with any particular study.

The first three, or basic, requirements relate as much to the study design as to the objective criteria for simple structure. Each factorial study for which there is to be an objectively defined simple structure should be so designed that the configuration of vectors satisfies these requirements. For the objective analytical criteria, on the other hand, these basic requirements form the essential framework. The first requirement parallels the concept of simple structure. The second requirement is necessary for the hyperplanes to be determinate. Consider, for example, a group of vectors for one hyperplane such that there was a two-dimensional space into which they only had small projections that could be interpreted as zero. The normal to the hyperplane could be located anywhere in this space and satisfy the first requirement. The location of the hyperplane would not be unique. In order for the location of the hyperplane to be definite it is necessary for the vectors in this hyperplane to have small projections into only one dimension, that of the normal to the hyperplane. The third requirement pertains most directly to the study design in the sense that there must be as many hyperplanes of vectors that satisfy the first two requirements as there are dimensions in the common-factor space. The study design should be such that the number of common factors extracted should be quite definite. When the third requirement is met by the study design, it is necessary, but probably not difficult, for the objective criteria to meet it also.

The types of freedom explicitly permitted in requirements four and five were selected because they touch on controversial, or possibly controversial, points. Factorial practice has been divided on the point of oblique versus orthogonal factors. It is the opinion of the author that in the present context maximum liberty should be permitted. Whenever it seems advisable, a restriction could be inserted to the effect that only orthogonal factors were permitted. This could be a function of the study being analyzed or of the opinion of the analyst. The case for complex measures has been previously

mentioned in this article. It is desirable for experimenters to be able to check in an objective fashion on hypotheses related to complex variables. Allowance for measures that have loadings on all factors is at variance with Thurstone's requirement (14, p. 335) that each row of the factorial matrix have at least one zero loading. In the opinion of the author this becomes an unnecessary restriction in case the basic requirements previously listed are met.

The last five, or operational, requirements relate to desirable aspects of objective criteria for simple structure. Requirement six could be met by the establishment of a range of projections, centering on zero, to be interpreted as negligible or zero projections. The limits for this range could be considered as generalized constants to be defined by the analyst on a priori grounds. A best fit of the data in some statistical sense as per requirement seven is certainly desirable. That this best fit should be unbiased, as per requirement eight, is also desirable. It is this requirement, however, that is likely to differentiate between an ideal objective criterion and various approximate ones. Requirements nine and ten are quite crucial, but at the same time may be the most difficult to satisfy. The statistical test of requirement nine is necessary for scientific acceptability, but it may be the last point to be solved for objective criteria for simple structure. The automatic computing procedures should be as economical as possible. It may be, however, that the computations for an ideal objective criterion will be so complex and extensive that such a criterion will be applied only to a few critical studies. Approximate criteria that involve simple computations might be adequate in many cases and would be highly desirable. Developments in high-speed computers, however, may influence the relative economies of the criteria.

Review of Previously Proposed Objective Criteria for Simple Structure

Turning next to an examination of proposed analytical definitions and procedures for a simple structure solution, Thurstone's equation for a simple structure will be considered first (11; 14, pp. 354-356). Thurstone makes the interesting proposal that his equation 28 is the equation for a simple structure.

$$\prod_{p=1}^r \left[\sum_{m=1}^r a_m \lambda_{mp} \right] = 0, \quad (1)$$

where p indicates simple structure factors, r is the number of factors, m indicates reference factors, a_m is a coordinate of a point on reference factor m , and λ_{mp} is the direction cosine on reference factor m of simple structure factor p . This equation states, in essence, that the product of the projections for each vector separately on the normals to the hyperplanes should be zero. This could be accomplished by the existence of at least one zero projection

for each vector. A least squares function for determining a best fit of the equation to data is suggested in Thurstone's equation 32.

$$\sum_{i=1}^n \prod_{p=1}^r \left[\sum_{m=1}^r a_{im} \lambda_{mp} \right]^2 = \phi, \quad (2)$$

where the notation is as above and j indicates tests. ϕ is to be minimized. No procedure is presented, however, for accomplishing this solution. Let us now consider this equation for simple structure in terms of our list of requirements for satisfactory objective criteria for simple structure. Zero projections are emphasized as per the first requirement. The second requirement is not necessarily satisfied, however, especially for batteries composed of very simple variables such that each vector might have a number of zero projections. Consider, for example, a battery composed of r groups of variables so that the vectors for each group form a separate cluster. In order for the vectors in each hyperplane to span a space of $(r - 1)$ dimensions, each hyperplane would have to pass through $(r - 1)$ of the clusters. This will, of necessity, result in each cluster being located in $(r - 1)$ of the hyperplanes. Thurstone's equation, however, may be satisfied by each cluster being located in only one hyperplane. Thus, each of the hyperplanes may be rotated so as to include only one cluster and not include vectors spanning an $(r - 1)$ -dimensional space. In this way Thurstone's equation does not satisfy our second requirement.

Thurstone's equation of a simple structure does involve as many simple structure factors as there are dimensions in the common-factor space and, thus, satisfies our third requirement. Both requirements on types of freedom permitted are met, except for permitting vectors with projections on all factors. Oblique factors may be involved. The variables may be complex up to the point of, but excluding, variables with projections on all factors. Among the operational requirements category, only the seventh and eighth requirements seem to be met. Thurstone has suggested a least squares function for the best fit of the equation to the data and this function seems to be unbiased in the sense of requirement eight.

Carroll (2) has proposed an analytic procedure that seems closely related to Thurstone's equation of a simple structure. In his development Carroll proposes "... that a satisfactory criterion for an approximation to simple structure is the minimization of the sums of cross-products (across factors) of *squares* of factor loadings." He obtains for each vector the products of each pair of projections on normals to the hyperplanes, sums these products for each vector and then over all vectors. Our first requirement that zero loadings are emphasized is satisfied. By employing products by pairs of projections Carroll circumvents the difficulty of Thurstone's equation in reference to our second requirement. Carroll's criterion is not necessarily satisfied by just one zero projection for each vector; thus, the solution tends

toward having each hyperplane determined in $(r - 1)$ dimensions. The third requirement is also satisfied in that a complete set of factors are considered. In the area of types of freedom permitted, either oblique or orthogonal factors may be used. There is a relation, however, between the use of complex variables and obtaining an unbiased fit to the data (requirements five and eight). Following a presentation of illustrative applications of his criterion, Carroll points out the biasing effects of complex tests and concludes that "These considerations lead to the conclusion that the present criterion will probably work best for well-designed factor studies where there are a large number of factorially pure tests and a relatively small number of factorially complex tests." (2, p. 33.) Requirements seven and ten are satisfied in that Carroll presents an objectively determined best fit and a procedure for accomplishing it. The procedure is laborious, but might be programmed for electronic computers. Requirements six and nine are not satisfied, but might be so by further developments and definitions. We conclude that Carroll's proposal is highly promising as an approximate method. It does satisfy the basic requirements, and tends to do so also for the types of freedom permitted, but it has some undesirable properties in the operational requirements area such that we agree with Carroll that his method is to be considered as yielding an approximation to simple structure.

Saunders (9, 10) has proposed a criterion for an approximation to simple structure involving the sum of fourth powers of factor loadings on orthogonal axes. Since it can be shown that Saunders' criterion is mathematically identical with Carroll's criterion discussed above when the orthogonal case is considered, we need not discuss Saunders' work extensively. In addition to an interestingly different and simpler computational procedure from that of Carroll, Saunders presents some comparisons of results from actual studies with results that were obtained from chance configurations of vectors. The results are quite promising.

Several other interesting recent publications involving closely related work to Carroll's development include articles by Ferguson (4), Neuhaus and Wrigley (7), and Pinzka and Saunders (8). Ferguson, starting from information theory, suggests using the sum of squares of products of factor loadings as a measure of parsimony, or lack of parsimony. Neuhaus and Wrigley in their quartimax method maximize the sum of the fourth powers of the factor loadings. A point of interest is their use of the Illiac (a high speed electronic computer). Pinzka and Saunders extended Saunders' solution to the oblique case. The discussion of the preceding two paragraphs applies directly to all three of these papers.

Thurstone in 1936 (12) proposed an analytic solution for simple structure involving a least squares solution of projections for a sub-group of variables for each hyperplane. The sub-group of variables was selected in terms of limiting sizes of projections on successive trials of an iterative procedure.

All of our requirements are met explicitly except two, three, and nine. The method used for selection of variables for the sub-groups allows the possibility that the essential dimensionality of the space spanned by the sub-group would be less than $(r - 1)$. By essential dimensionality we mean the number of dimensions in which some vectors for the sub-group have projections that would not be interpreted as zero (that is, less than the stated limits on size of projections used in selecting the variables). In our proposal, to be discussed later, an objective procedure is indicated which will circumvent our objection to this method of Thurstone. For Thurstone's method as he proposed it, we feel that failure to guarantee that the sub-group spanned an $(r - 1)$ -dimensional space was a serious drawback which would make the method unacceptable. Requirement three could be met for each study by a succession of solutions, each involving location of a single hyperplane, until as many distinct hyperplanes were found as there were dimensions in the common factor space of the study. There is no guarantee, however, that all such hyperplanes could be found.

A variant of Thurstone's preceding procedure was presented by Horst (6), in which he maximized the ratio of the sum of squares of significant projections to the sum of squares of all projections for each hyperplane. This is mathematically equivalent to minimizing the ratio of the sum of squares of the non-significant projections to the sum of squares of all projections. Again the difference between significant and non-significant projections was made in practice on size of projection in successive trials. Comments on this method are identical with those on the preceding method.

Tucker (16, 17) proposed non-analytical procedures making use of graphs and judgment of the analyst designed to insure that the sub-groups did span spaces of $(r - 1)$ dimensions. In that these procedures involve subjective judgments in the process of analysis they will not be evaluated here. Their importance here is that they did attempt to solve one of the more important problems in the determination of simple structure hyperplanes. It is possible by continually reducing the sub-group of variables to obtain a sub-group that will have non-significant projections in one direction. The problem is to guarantee that such a sub-group does have some significant projections in all directions orthogonal to the one for which the projections are non-significant.

Thurstone has recently proposed a still different type of objective procedure (15) in which a minimum weighted sum of projections is obtained. The weights are related to the projections by an arbitrary step function so as to emphasize near zero projections. This is a single-plane method in that one hyperplane is determined at a time. Although only projections on successive trial normals are used, the distinction between significant and non-significant projections is not a sharp break but rather a transition dependent on lower weights for projections of intermediate size. This will

increase the chance of involving variables spanning $(r - 1)$ dimensions in the determination of the hyperplane. In that the range of projections that receive finite weights is broad there is a chance that the solution could be biased in the sense used in our requirement eight. Vectors with significant projections on the normal could influence the location of the normal and thus produce a non-zero mean of non-significant projections even when the variance of the non-significant projections was low. We conclude that this latest objective method should be classified as an approximate procedure. It may be a very useful procedure, however, since the computations are quite simple and the results presented by Thurstone indicate good approximations to the desired results.

Definition of Simple Structure by Linear Constellations and Vector Masses

In the objective definition of simple structure proposed here a concept of linear constellations is employed. Consider the left half of Figure 1. This

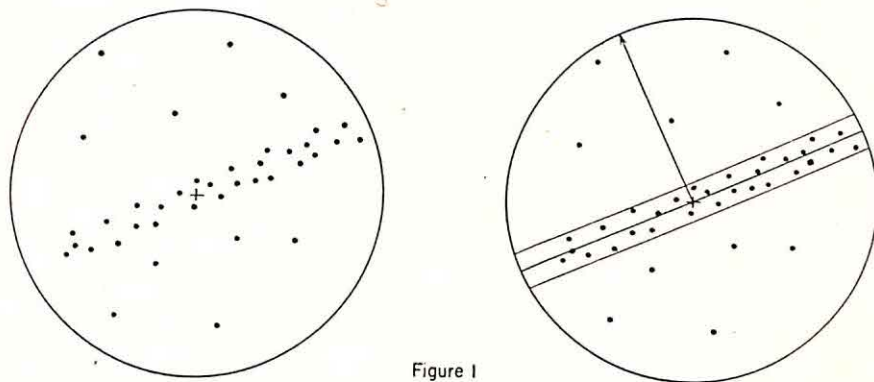


Figure 1

is a two-dimensional view of a factorial geometric model. Other dimensions are orthogonal to the plane of the figure. Each dot is the projection of the terminus of a vector representing a variable included in the battery being analyzed. It is postulated for Figure 1 that the battery of variables is such that the vectors might appear in a band such as is shown. If a direction is chosen orthogonal to this band, the vectors represented by the dots concentrated in this band will have small projections. In terms of a parametric explanation of the variances of the variables there will be a corresponding low dependence of these variables on a parameter corresponding to the direction orthogonal to the band. Such concentrations of vectors into linear spaces which include the origin may be termed linear constellations.

At the right of Figure 1, a line through the band of points and two bounding lines have been drawn to indicate the space of the linear constellation and the limits for projections outside this space. In general, linear constellations may be of any dimensionality less than that of the common-

factor space for the entire battery of variables. When the constellation contains only one dimension it would be called a cluster. This one dimension would represent a single parameter and could be interpreted. In case the linear constellation has as many dimensions less one as the common-factor space, the constellation may be designated by the one dimension orthogonal to the constellation. This normal can be used to indicate a parameter not involved in the constellation. The projections of the vectors on this normal will indicate the extent of dependence of the observed variables on this parameter. A simple structure is interpreted in the present context as a set of these linear constellations, the number of constellations in the set being equal to the dimensionality of the common-factor space.

The problem of defining simple structure is now transformed into that of explicitly defining linear constellation with dimensionality one less than the common-factor space. Let these constellations be termed linear constellations of dimensionality ($r - 1$). Steps in the operational definition of such a constellation include the following:

1. Appended to and equally on both sides of any and every hyperplane in the common-factor space is a marginal space of some defined and limited width.
2. Any vector located entirely within a hyperplane and its marginal space shall be considered as contained in the hyperplane.
3. The number of vectors contained in a hyperplane shall be termed the vector mass of the hyperplane.
4. A maximum vector mass for a hyperplane occurs when rotation of the hyperplane in any direction results in a decrease in the vector mass before any subsequent increase in vector mass. (It is to be noted that with a finite number of vectors the location of the hyperplane for a maximum vector mass will not be unique. Small rotations of the plane may not result in a change in the vector mass.)
5. Those vectors contained in a hyperplane when the vector mass is maximum constitute a linear constellation of dimensionality ($r - 1$) and the hyperplane will be termed the space of the linear constellation.

Definition of a simple structure adds the following step:

6. A simple structure is constituted by the hyperplanes for a set of r linear constellations of dimensionality ($r - 1$).

A comparison of the foregoing definition with our requirements indicates that all requirements are met with the exception of number nine, relating to a statistical test, and number ten, concerning an automatic computational procedure. Emphasis is placed on a maximum concentration of vectors along the hyperplanes (requirement one). A maximum vector mass occurs only when the vectors contained in the hyperplane span a space of ($r - 1$) dimensions, for otherwise a rotation would result in an increase in the vector mass (requirement two). In order to clarify this point, consider a group of vectors that are contained in a space of ($r - 2$) dimensions and an appended space

of the defined radial width in the other two dimensions. In a three-dimensional factorial space such a group of vectors would form a cluster around a single direction. This group of vectors is contained in any hyperplane whose normal lies in the two-dimensional plane orthogonal to the given $(r - 2)$ -dimensional space containing the group of vectors. Any hyperplane that contains just this group of vectors, therefore, may be rotated without loss of this group of vectors and may be made to contain one or more vectors not contained in the given $(r - 2)$ -dimensional space. This step depends on the existence of vectors not contained in the $(r - 2)$ -dimensional space, but such vectors must exist for the common-factor space to be of r dimensions. Thus, the vector mass of the hyperplane can be increased before any decrease occurs, and the original position of the hyperplane did not possess a maximum vector mass. This argument can be extended to vector groups contained in spaces of $(r - 3)$ or fewer dimensions. In consequence, a maximum vector mass occurs only when the vectors contained in the hyperplane are *not* contained also in a space of $(r - 2)$ or fewer dimensions; that is, the vectors contained in such a hyperplane must span a space of $(r - 1)$ dimensions.

The simple structure is defined in step six as being constituted by r hyperplanes, which is the dimensionality of the common-factor space (requirement three). No limitations are placed as to oblique or orthogonal factors or as to complexity of a minority of the tests (requirements four and five). A defined limit for projections of vectors to be contained in the hyperplane is indicated in our definitions one and two (requirement six). The linear constellations are objectively defined by maximum vector mass (requirement seven). This definition is unbiased since the marginal space of definition one is appended equally to both sides of the hyperplane (requirement eight).

It is hoped that one could derive a statistical test such as is indicated in requirement nine. Such a development would make a definite contribution to the field of factor analysis. At present, however, this requirement for a satisfactory criterion of simple structure has not been satisfied.

Computing Procedure for Linear Constellations

An automatic method for searching for linear constellations, as per requirement ten, has been developed and tried out. The labor of computations is quite great, but within bounds for automatic computing machinery. One trial has involved a run on an IBM Card Programmed Calculator. In addition a careful check has been made in detail on the feasibility of performing the computations on the IBM Type 701 Electronic Computer. This machine could perform the required computations on an automatic basis within feasible time, such as 10 minutes for 50 variables in 10 dimensions for each linear constellation.

It is of interest that the method finally adopted as feasible is a combination of two methods neither of which is feasible. The first of these methods

Step 1: List a matrix F_o for a selected sub-group of tests (see Table 2). For the first cycle the sub-group might be taken as those tests that have low correlations with some particular test. In experimental applications of this method, initial sub-groups were usually taken to contain approximately half of the tests in the battery. It was found that each of the linear constellations resulted from several different initial sub-groups. Enough different initial sub-groups were used for the study employed in the example to be able to find three distinct linear constellations. Two points of general concern are the recognition of duplicating results and being able to find all existing linear constellations. Any duplication can be readily detected by comparisons of the solutions and may be eliminated by discarding results from one or more initial sub-groups. The problem of selection of sub-groups so as to be able to find all linear constellations is much more difficult. After a number of constellations are found, a vector might be set orthogonal to them and tests selected that have low projections on this vector. Another possibility is to first employ a method such as Carroll's (2), or Saunders' (9, 10) and to establish initial sub-groups of tests with low projections on each of the factors so determined.

The initial sub-group in the example contains tests 3, 4, and 1. For the second and subsequent cycles the sub-groups are given by the preceding cycle.

Step 2: Compute the matrix P_o (see Table 2).

$$P_o = F_o' F_o . \quad (3)$$

Step 3: Compute the two smallest characteristic vectors of P_o (see Table 2). These are the characteristic vectors corresponding to the two smallest characteristic roots of P_o . The smallest vector is C_1 and the next to smallest vector is C_2 . Each of these vectors is to be a unit vector (have sum of squares of entries equal to unity). The matrix containing these two vectors is labeled Λ in Table 2.

Step 4: Compute the matrix projections, V , of all the tests on the two smallest characteristic vectors (see Table 2).

$$V = F\Lambda . \quad (4)$$

Step 5: Survey the space of the two smallest characteristic vectors for the radial band of specified width which includes the largest number of test vectors. The concept involved is illustrated in Figure 2. A plot between projections of the tests on C_1 and C_2 is shown on the left. The dots for our trial sub-group of tests 3, 4, and 1 are located near the origin. Centered on C_1 and indicated by short lines outside the circle are eleven directions separated by 9° . The line with an arrow is pointing in the direction of -36° . Orthogonal to this trial normal is a line for the tentative linear subspace and two limit lines. The trial normal was also placed in each of the other

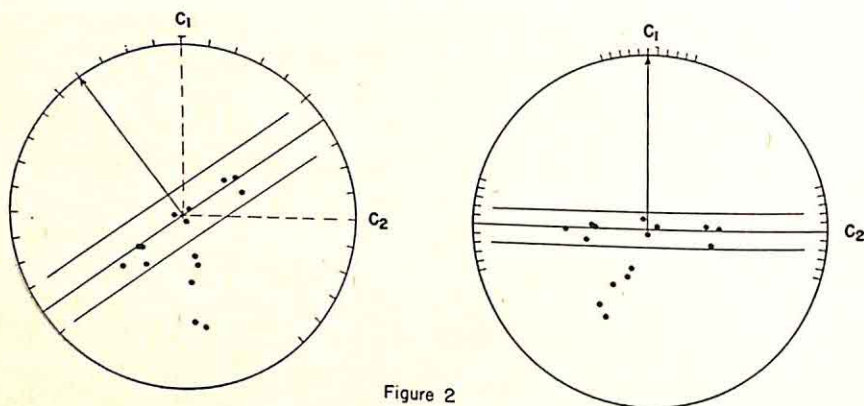


Figure 2

ten selected directions. The short lines inside the circle indicate the corresponding locations of the linear subspace. For each of the set of directions in this survey, a count was made of the number of points between the corresponding limit lines. For the direction in which the lines are drawn, ten of the dots lie in the space between the two limits. This would also be true for the trial normal placed at -45° . All other nine directions have fewer points in the space between the limits. The ten tests for the dots lying between the limit lines were selected for the next sub-group for the next cycle.

In practice, the plots in Figure 2 would not be made since the operations can be performed by computing steps illustrated in Table 3 and outlined below:

a. Define a transformation matrix U for the set of survey vectors to be employed. This matrix will contain the direction cosines of the survey vectors in terms of C_1 and C_2 . Two such matrices are given in Table 5, a coarse survey set with 9° steps and a fine survey set with 3° steps. The coarse survey set was used in Table 3.

b. Find the projections of all tests on each of the survey vectors. These projections are contained in the matrix V_s of Table 3:

$$V_s = VU. \quad (5)$$

In this table, the test numbers for the sub-group of tests are double-starred.

c. Establish limits for projections to be considered as negligible and count the number of projections in each column of V_s within these limits. In the survey given in Table 3, limits of .15 and $-.15$ were used. All projections within these limits are starred, and the number of such projections in each column is given at the bottom of the table.

d. Choose the column of projections in V_s with the largest number of negligible projections. In case of a tie in this count between two columns, choose the column for the smallest angular deviation from C_1 . In the example in Table 3, both columns -45° and -36° have counts of ten negligible projections. According to our rule to choose the column with the smallest angular deviation from C_1 we chose the -36° column. A possible minor problem that may arise is when there is a tie between a column with a positive angular deviation from C_1 and a column with an equal negative angular deviation from C_1 . In this case, an arbitrary decision might be made to choose the column with the positive angular deviation from C_1 .

e. Select the tests with negligible projections on the chosen survey vector of step d

as the sub-group for the next cycle of computations. In the example of Table 3, the tests with starred projections in column -36° were selected as the revised sub-group for the next cycle of computations.

It is anticipated that a rather coarse survey will be used during initial cycles for one of the factors, or linear constellations. When the sub-group of

TABLE 4

Fine Survey for Revised Sub-group												
Matrix $V_{62} = V_2 U_2^T$												
Projections on Survey Vectors												
Test No.	-15°	-12°	-9°	-6°	-3°	0°	3°	6°	9°	12°	15°	
10**	-17	-15	-14	-12	-10	-08*	-06*	-04*	-02*	00*	02*	
2**	-09*	-07*	-05*	-03*	-01*	01*	03*	05*	07*	09*	11	
5**	-06*	-04*	-02*	-01*	01*	03*	05*	07*	08*	10	11	
3**	01*	01*	01*	01*	02*	02*	03*	03*	03*	03*	03*	
4**	-03*	-03*	-03*	-03*	-03*	-03*	-03*	-03*	-03*	-03*	-03*	
1**	07*	07*	06*	06*	06*	06*	06*	06*	05*	05*	05*	
8**	09*	07*	06*	06*	06*	06*	06*	06*	04*	04*	04*	
7**	10	09*	07*	06*	06*	06*	06*	06*	05*	05*	05*	
9**	02*	00*	-01*	-03*	-05*	-07*	-09*	-11	-13	-14	-16	
6**	11	09*	07*	04*	01*	01*	03*	06*	09*	-11	-13	
15	-23	-24	-25	-26	-26	-27	-28	-28	-28	-29	-29	
14	-26	-27	-28	-30	-31	-32	-33	-34	-35	-36	-36	
17	-35	-36	-38	-40	-42	-43	-44	-46	-47	-48	-49	
11	-42	-44	-46	-47	-49	-50	-51	-53	-54	-55	-55	
18	-19	-20	-20	-21	-22	-22	-22	-23	-23	-23	-24	
n												
$ V < 10$	7	9	9	9	9	10	10	9	9	7	6	

Coarse Survey for Initial Sub-group												
Matrix $V_{61} = V_1 U_1^T$												
Projections on Survey Vectors												
Test No.	-45°	-36°	-27°	-18°	-9°	0°	9°	18°	27°	36°	45°	
10	-14*	-09*	-03*	03*	08*	14*	19	24	28	31	34	
2	-06*	00*	07*	12*	18	23	28	31	34	37	38	
5	-02*	03*	08*	13*	17	21	25	27	29	31	32	
3**	00*	01*	01*	02*	02*	03*	03*	04*	04*	04*	04*	
4**	-04*	-04*	-04*	-04*	-04*	-04*	-04*	-03*	-03*	-02*	-01*	
1**	04*	03*	02*	01*	00*	-01*	-02*	-02*	-03*	-03*	-04*	
8	02*	-03*	-07*	-12*	-16	-20	-23	-26	-28	-30	-31	
7	04*	-01*	-06*	-11*	-16	-20	-24	-27	-30	-32	-33	
9	-05*	-11*	-16	-21	-25	-29	-32	-34	-36	-36	-36	
6	03*	-04*	-12*	-19	-25	-31	-36	-40	-43	-46	-47	
15	-27	-29	-30	-30	-29	-27	-25	-22	-18	-14*	-11*	
14	-31	-35	-37	-39	-39	-38	-35	-32	-29	-24	-24	
17	-42	-46	-49	-52	-53	-52	-50	-47	-43	-38	-32	
11	-48	-52	-55	-56	-57	-55	-52	-48	-43	-37	-30	
18	-22	-24	-25	-25	-25	-24	-23	-21	-18	-15	-12*	
n												
$ V < 15$	10	10	9	8	4	4	3	3	3	3	5	

TABLE 5

Coarse Survey:												
Matrix U_c												
Degrees Deviation from Smallest Characteristic Vector												
	-45	-36	-27	-18	-9	0	9	18	27	36	45	
C1	.70711	.80902	.89101	.95106	.98769	1.00000	.98769	.95106	.89101	.80902	.70711	
C2	-.70711	-.58779	-.45399	-.30902	-.15643	.00000	.15643	.30902	.45399	.58779	.70711	

Fine Survey:												
Matrix U_f												
Degrees Deviation from Smallest Characteristic Vector												
	-15	-12	-9	-6	-3	0	3	6	9	12	15	
C1	.96593	.97815	.98769	.99452	.99863	1.00000	.99863	.99452	.97815	.96593		
C2	-.25882	-.20791	-.15643	-.10453	-.05234	.00000	.05234	.10453	.15643	.20791	.25882	

Survey Matrices

tests is not altered by a cycle of computations with a coarse survey, a finer survey may be employed. This finer survey would involve smaller angular steps for the survey vectors and narrower limits for negligible projections. Such a fine survey is illustrated at the right of Figure 2 for the illustrative problem and is given in Table 4. Three-degree steps were used in the second U matrix of Table 5, and limits of .10 and -.10 were used. In this case the sub-group for the cycle was composed of the tests indicated in Table 3 for the first cycle. These test numbers are double starred in Table 4. A series of fine surveys might be required before there is no change in the sub-group. When there is no change in the sub-group as illustrated in Table 4 (the 0° is the chosen column), the smallest characteristic vector, C_1 , is the normal to the desired hyperplane of the linear constellation, or factor.

This method has been tried on the illustrative example to determine three linear constellations by starting from different trial subgroups. These trial sub-groups were determined in this case as variables which had low correlations with selected variables. An alternative approach would be to apply one of the approximate solutions such as Carroll's (2) or Saunders' (9) and to pick variables with low projections on the resulting factors. In any case before the computations are initiated by the present method it is necessary to select a number of initial trial sub-groups and to define the two limits to be used in the coarse and fine surveys. Otherwise, the computations are completely automatic until a stable solution is obtained for each initial sub-group. At the end it will be necessary to compare the results from the several initial sub-groups and eliminate any duplications. In case the number of linear constellations discovered is less than the number of dimensions in the common-factor space, new initial sub-groups might be tried. Thus, this computing procedure is not so sure to find all of the linear constellations that are indicated in the definitions and may not satisfy our third requirement. It should yield satisfactory results, however, for those well-designed studies in which the vectors are concentrated along all hyperplanes.

REFERENCES

1. Brigham, C. C. A study of error. New York: College Entrance Examination Board, 1932.
2. Carroll, J. B. An analytic solution for approximating simple structure in factor analysis. *Psychometrika*, 1953, 18, 23-38.
3. Cattell, R. B. Factor analysis. New York: Harper, 1952.
4. Ferguson, G. A. The concept of parsimony in factor analysis. *Psychometrika*, 1954, 19, 281-290.
5. Holzinger, K. J. and Harman, H. H. Factor analysis. Chicago: Univ. Chicago Press, 1941.
6. Horst, P. A non-graphical method for transforming an arbitrary factor matrix into a simple structure factor matrix. *Psychometrika*, 1941, 6, 79-100.
7. Neuhaus, J. O. and Wrigley, C. The quartimax method. *Brit. J. stat. Psych.*, 1954, 7, 81-91.

8. Pinzka, C. and Saunders, D. R. Analytic rotation to simple structure, II: Extension to an oblique solution. Educational Testing Service Bulletin, RB-54-31, 1954. (Multilithed)
9. Saunders, D. R. An analytical method for rotation to orthogonal simple structure. *Amer. Psychologist*, 1953, 8, 428. (Abstract)
10. Saunders, D. R. An analytical method for rotation to orthogonal simple structure. Educational Testing Service Research Bulletin RB-53-10 1953. (Multilithed)
11. Thurstone, L. L. The vectors of mind. Chicago: Univ. Chicago Press, 1935.
12. Thurstone, L. L. The bounding hyperplanes of a configuration of traits. *Psychometrika*, 1936, 1, 61-68.
13. Thurstone, L. L. An experimental study of simple structure. *Psychometrika*, 1940, 5, 153-168.
14. Thurstone, L. L. Multiple-factor analysis. Chicago: Univ. Chicago Press, 1947.
15. Thurstone, L. L. An analytical method for simple structure. *Psychometrika*, 1954, 19, 173-182.
16. Tucker, L. R. A rotational method based upon the mean principal axis of a sub-group of tests. *Psychol. Bull.*, 1940, 37, 578. (Abstract)
17. Tucker, L. R. A semi-analytical method of factor rotation to simple structure. *Psychometrika*, 1944, 9, 43-68.
18. Tucker, L. R. An objective determination of simple structure in factor analysis. *Amer. Psychologist*, 1953, 8, 448. (Abstract)
19. Vernon, P. E. The structure of human abilities. London: Methuen & Co., Ltd., 1950. (New York: Wiley.)

Manuscript received 8/20/54

Revised manuscript received 11/15/54

F-TEST BIAS FOR EXPERIMENTAL DESIGNS IN EDUCATIONAL RESEARCH

NEIL GOURLAY

UNIVERSITY OF BIRMINGHAM

Reference is made to Neyman's study of *F*-test bias for the randomized blocks and Latin square designs employed in agriculture, and some account is given of later statistical developments which sprang from his work—in particular, the classification of model-types and the technique of variance component analysis. It is claimed that there is a need to carry out an examination of *F*-test bias for experimental designs in education and psychology which will utilize the method and, where appropriate, the known results of this new branch of variance analysis. In the present paper, such an investigation is carried out for designs which may be regarded as derivatives of the agricultural randomized blocks design. In a paper to follow, a similar investigation will be carried out for experimental designs of the Latin square type.

I. Introduction

F-test bias may be said to exist for a given experimental situation if, when the null hypothesis is valid, frequent replication of the experiment provides a distribution of *F*-values which does not conform in some way (within the limits of sampling error) to the corresponding theoretical *F*-distribution. When bias exists, it is important for the investigator to know whether the *F*-test (the null hypothesis being valid) gives a larger or smaller proportion of significant *F*-ratios than is warranted by the theoretical distribution.

The possibility of *F*-test bias for certain experimental designs first became a topic of major statistical interest with Neyman's paper (14) in 1935. Neyman confined his inquiry to the randomized blocks and Latin square designs, which Fisher had developed; these designs had become the mainstay of agricultural experimentation. In both cases, he pointed out that "the conditions under which the application of the *z*-distribution is legitimate are not strictly satisfied" and went on to show that "in the case of the randomized blocks the position is somewhat more favorable to the *z*-test, while in the case of the Latin square this test seems to be biased, showing the tendency to discover differentiation when it does not exist."

Neyman's conclusions met at first with considerable opposition, but as Kendall (10, p. 214) points out, the controversy arose mainly from a failure to realize that Neyman was dealing with a different hypothesis from that usually tested. Thus, Fisher was concerned with the hypothesis that for each plot in the experimental field the treatments had the same effect. Neyman, on the other hand, stressed the possibility of interactions between plots and treatments and considered the more general hypothesis that the

mean effects of treatments over all plots involved in the experiment were the same. In 1937, Welch (20) clearly distinguished between the two hypotheses and went on to show that the z -distribution furnished an approximate test of the Fisher hypothesis both for the randomized blocks and Latin square designs. His findings did not, of course, invalidate in any way Neyman's analysis.

With regard to the validity of Neyman's analysis, it might be pointed out that while Neyman insists that the correction for fertility of a plot may vary from treatment to treatment, he regards the fertility corrections for blocks (in the case of randomized blocks) and for rows and columns (in the case of the Latin square) as being the same for all treatments. This assumption does not appear justifiable and, if not made, Neyman's results might be modified considerably.

Important as Neyman's investigation was for the randomized blocks and Latin square techniques, later years showed that his work was to exert a more widespread influence. Not only did he provide a method of analysis for detecting bias in special cases, but he aroused interest in the problem of bias generally. In addition, it was quickly realized that his method of analysis could also be employed to find estimates for the components of variance in any given experimental set-up. There arose a new branch of variance theory known generally as variance component analysis [cf. Crump (5)]. Further, the new approach made statisticians much more cognizant of the types of problem with which they had to deal: attention became focused on the types of mathematical model which they applied to different situations, and which formed the basis of their statistical analyses [cf. Eisenhart (8) and Crump (5)].

When we turn to consider the field of educational and psychological research it would appear that the randomized blocks and Latin square techniques were absorbed into this field without any noticeable recognition of the possible relevance of Neyman's findings. Recently, however, McNemar (13) has pointed out to users of the Latin square in psychology that they have ignored the fundamental assumption that all interactions are zero; and after stating, without giving or quoting any analysis in support, that failure to satisfy this assumption will lead to too many significant F 's, he concludes that the Latin square technique is seldom appropriate and that "it is defensible only in those rare cases where one has sound a priori reasons for believing that the interactions are zero." Also it would be true to say that little reference, if any, has been made to the results of the other investigators who have continued the work that Neyman began. Two reasons might be offered in explanation. First, Neyman and many of the others were concerned with agricultural research and, consequently, were dealing with experimental situations which do not normally exist in education and

psychology. Secondly, many of the articles are of too recent origin for their results to appear to any great extent in the textbooks and research publications belonging to the latter field.

There is obviously a need to carry out an examination of F -test bias for experimental designs employed in education and psychology. Some such research has already been reported but it requires to be supplemented. The method and, where appropriate, the known results of variance component analysis should be utilized. In the present paper such an investigation is carried out for designs which may be regarded as derivatives of the agricultural randomized blocks design. In a second paper, the same will be done for those designs of which the agricultural Latin square is the prototype.

II. Method

For a valid (i.e., unbiased) F -test, the two variances involved in the F -ratio must be independent unbiased estimates—based on the stated numbers of degrees of freedom—of the same normal population variance. Test bias arises when the variances of the F -test fail to satisfy this set of conditions in one or more respects. It follows that, in order to detect bias, it is sufficient to examine the data—by simple inspection or by statistical analysis—for any failure to comply with the conditions of normality, homogeneity of variance, independence of estimates, etc.

It is not, however, sufficient to know that bias exists. An investigator also wants to know (a) the direction of the bias and (b) its magnitude; or, at least, he wants some indication of the answer to both these questions. For convenience, we will define an F -test to be positively or negatively biased, if, in the case where the null hypothesis being tested is correct, the test produces a larger or smaller proportion, respectively, of significant F -ratios than is warranted by the F -distribution.

In this paper considerable use will be made of Neyman's procedure (i.e., variance component analysis) as a method of detecting bias and of indicating its direction and magnitude. (The other and perhaps more common use of this form of analysis to obtain estimates of variance components will be involved only incidentally.) The method consists simply of taking the mathematical model which applies to the experimental situation and deriving analytically the expected values of the variances involved in the F -test. Then, in the case where the null hypothesis holds, the expected value of the "treatments" variance will be equal in magnitude to that of the "error" variance *if no bias is present*. When the two expected values are unequal, positive or negative bias is suggested according as the first variance is greater or less than the second. Also some measure of the magnitude of the bias is provided by the amount the ratio of the two expected values differs from unity.

For ease of exposition, we shall refer to this ratio as the B -ratio. Thus positive or negative bias is suggested by B -ratios greater or less than unity, respectively.

The method has several limitations:

(i) A B -ratio of unity is a necessary but not a sufficient condition for zero bias (the empirical F -distribution may have the same mean as the theoretical F -distribution but may differ from the latter in respect of standard deviation or any other moment). Consequently, it frequently happens that bias is present although the B -ratio is unity. Several instances of this appear in the present study.

(ii) As might be expected from (i), the value of the B -ratio is by no means always a certain criterion of the direction of bias; but it is probably true to say that it gives a correct indication of bias direction for most types of analyses where its value is other than unity.

(iii) In the same way, the deviation of the B -ratio from unity is only a very rough indication of the magnitude of bias. Obviously account must also be taken of the numbers of degrees of freedom involved in the F -test. A very tentative procedure for doing this will be suggested later.

Where Neyman's procedure is inadequate, other methods of bias analysis are required, and if these, because of the mathematical difficulty, are not easy to devise, empirical methods must be adopted. No such empirical studies are attempted in this paper, but reference is made to one or two studies of this type.

III. Models

All the models involved in this paper are linear models applying to a two-way classification. It may be helpful to the reader if, before proceeding to the main discussion, he is given an account of some of the models of this type to be found in the literature and is shown how the models of the present paper are related to them.

Eisenhart (8) distinguishes three types of models. In describing these we shall follow Crump (5) and adopt a broader interpretation than that chosen by Eisenhart.

Model I (*Fixed variate model*)

This may be written

$$X_{rst} = \mu + A_r + B_s + I_{rs} + \epsilon_{rst} \quad \left\{ \begin{array}{l} r = 1, \dots, p \\ s = 1, \dots, q \\ t = 1, \dots, n \end{array} \right. \quad (1)$$

where X_{rst} represents the t th observation in the subclass (r, s) , μ is the general mean, A_r and B_s are the main effects for the corresponding column and row,

respectively, I_{rs} denotes the interaction effect for the r th column and s th row, and ϵ_{rst} is the random error for the observation.

The population of A 's, B 's and I 's are all finite (of zero mean) and are exhausted in the given $p \times q$ classification, but the population of ϵ 's is continuous with a normal probability distribution of variance σ_ϵ^2 .

The expected values of the mean squares involved in the analysis of variance of the data are as follows:

	d. f.	Expected Value of Mean Square
Columns	$p - 1$	$\sigma_\epsilon^2 + qn \sum_r A_r^2 / (p - 1)$
Rows	$q - 1$	$\sigma_\epsilon^2 + pn \sum_s B_s^2 / (q - 1)$
Interaction	$(p - 1)(q - 1)$	$\sigma_\epsilon^2 + n \sum_r \sum_s I_{rs}^2 / (p - 1)(q - 1)$
Residual	$pq(n - 1)$	σ_ϵ^2

It will be seen that the null hypotheses (i) $A_r = 0$ ($r = 1, \dots, p$), (ii) $B_s = 0$ ($s = 1, \dots, q$), (iii) $I_{rs} = 0$ ($r = 1, \dots, p$; $s = 1, \dots, q$) are tested by examining the significance of the F -ratios of *columns*, *rows*, and *interaction*, respectively, with respect to *residual*. Normally when *interaction* is significant, the investigator is not interested in making the test for *columns* and *rows* although there is no theoretical objection to his doing so. (Eisenhart actually restricts Model I to the case of zero interaction by making his second assumption of additivity).

Model II (Random variate model)

This may be written

$$X_{rst} = \mu + \alpha_r + \beta_s + \eta_{rs} + \epsilon_{rst} \quad \left\{ \begin{array}{l} r = 1, \dots, p \\ s = 1, \dots, q \\ t = 1, \dots, n \end{array} \right\}, \quad (2)$$

where the terms may be described as for the corresponding members of Model I but, in this case, the p α -values, q β -values and pq η -values are random samples from normal distributions of zero mean and of variance σ_α^2 , σ_β^2 , and σ_η^2 , respectively.

The expected values for the mean squares in the variance analysis are:

	d. f.	Expected Value of Mean Square
Columns	$p - 1$	$\sigma_\epsilon^2 + n\sigma_\eta^2 + nq\sigma_\alpha^2$
Rows	$q - 1$	$\sigma_\epsilon^2 + n\sigma_\eta^2 + np\sigma_\beta^2$
Interaction	$(p - 1)(q - 1)$	$\sigma_\epsilon^2 + n\sigma_\eta^2$
Residual	$pq(n - 1)$	σ_ϵ^2

The null hypothesis $\sigma_{\eta}^2 = 0$ is tested by testing *interaction* against *residual*. The hypotheses $\sigma_{\alpha}^2 = 0$, $\sigma_{\beta}^2 = 0$ are tested by testing *columns* and *rows* respectively against *interaction* or, where it is known *a priori* that interaction is zero, against *total* residual of $(pqn - p - q + 1)$ d. f. (It will be seen from the table that when $\sigma_{\eta}^2 = 0$, *columns* and *rows* can be tested against either *interaction* or *residual*. The latter provides the more precise test but a further increase in precision is obtained if the sums of squares for *interaction* and *residual* are combined to form an estimate of σ_{ϵ}^2 based on $(pqn - p - q + 1)$ d. f. and the tests of *columns* and *rows* made against this *total* residual).

Mixed Model

This takes the form

$$X_{rst} = \mu + A_r + \beta_s + \eta_{rs} + \epsilon_{rst} \quad \left\{ \begin{array}{l} r = 1, \dots, p \\ s = 1, \dots, q \\ t = 1, \dots, n \end{array} \right\}, \quad (3)$$

where the population of A -values is finite (of size p) but the β - and η -values are random samples from infinite populations. The expected values of the mean squares now read:

	d. f.	Expected Value of Mean Square
Columns	$p - 1$	$\sigma_{\epsilon}^2 + n\sigma_{\eta}^2 + nq \sum_r A_r^2 / (p - 1)$
Rows	$q - 1$	$\sigma_{\epsilon}^2 + n\sigma_{\eta}^2 + np \sigma_{\beta}^2$
Interaction	$(p - 1)(q - 1)$	$\sigma_{\epsilon}^2 + n\sigma_{\eta}^2$
Residual	$pq(n - 1)$	σ_{ϵ}^2

Tests of hypotheses are made as for Model II.

Useful as the above classification is, it fails to cover many of the cases which occur in practice. Thus, the models of Fisher and Neyman for the agricultural randomized blocks design belong to quite a distinct class. A more extensive classification has been proposed by Tukey [cf. Crump (5)].

In the present paper, the models studied may be regarded as modified versions of Eisenhart's Mixed Model: there is only one exception, which is a special case of Model I.

The basic equation for these modified versions may be written

$$X_{rst} = \mu + A_r + \beta_s + \eta_{rs} + \xi_{rs} + \epsilon_{rst} \quad \left\{ \begin{array}{l} r = 1, \dots, p \\ s = 1, \dots, q \\ t = 1, \dots, n_{rs} \end{array} \right\}. \quad (4)$$

The main difference between this and Eisenhart's Mixed Model is the additional random error term ξ_{rs} , common to all observations in the subclass (r, s) . As will be seen later, this term differs from the η -term in that the ξ -values are usually regarded as independent (uncorrelated) while the η -values may be correlated and, what is more, show heterogeneity of correlation (between columns).

The next section, dealing with the investigation of bias for these models, falls conveniently into three parts:

1. Equal numbers in subclasses, i.e., $n_{rs} = \text{const.} = n$, say.
 2. Numbers in subclasses unequal but proportional, i.e., $n_{rs} = Na_r b_s$, where N is total number of cases sampled and a_1, \dots, a_p and b_1, \dots, b_q are the proportions of cases in columns and rows, respectively ($\sum_r a_r = 1 = \sum_s b_s$).
 3. Numbers in subclasses unequal and disproportionate.
- Types of bias common to all three cases are discussed in the first part.

IV. Investigation and Results

1. Equal Numbers in Subclasses ($n_{rs} = n$)

It will make the discussion more concrete and less theoretical if we speak in terms of a methods experiment replicated in a random sample of schools. Lindquist (12) gives an excellent account of the experimental design and statistical analysis required for this type of experiment. The main F -test in the analysis is that of the *methods* variance against the *interaction* variance. The hypothesis tested is that the methods have the same mean effect over the *total* population of schools.

The interaction term of the analysis not only contains sampling error (measured by the variance *within classes*) but it may, and usually does, contain two other elements:

- (i) real interaction between methods and schools;
- (ii) group errors, i.e., errors which apply to the experimental groups as wholes and which are produced by factors other than method and school differences, e.g., teacher differences.

It will be seen that the model for this type of design is a version of the special mixed model mentioned at the end of the last section, namely,

$$X_{rst} = \mu + A_r + \beta_s + \eta_{rs} + \xi_{rs} + \epsilon_{rst} \quad \left\{ \begin{array}{l} r = 1, \dots, p \\ s = 1, \dots, q \\ t = 1, \dots, n \end{array} \right\}, \quad (5)$$

where μ is the general mean and the A, β, η, ξ and ϵ represent the effects due to methods, schools, interaction, group error, and sampling error, respectively.

As usual $\sum_{r=1}^p A_r = 0$. ξ_{rs} and ϵ_{rst} are random, the parent populations

being assumed to be normal, of zero mean and of variance σ_{ξ}^2 and σ_{ϵ}^2 , respectively. β_s is usually defined so that $(\mu + \beta_s)$ is the mean for the s th school over the p methods (and the total population of ξ - and ϵ -values). But it might be more instructive if we here take $(\mu + \beta_s)$ to be the mean of the s th school over a population of methods which includes the p methods under consideration. The parent population of β -values will be assumed to be infinite and of variance σ_{β}^2 (the mean of course is zero).

Since some of the methods within the total population of methods will normally resemble one another more than they do the others, the interaction terms for these methods (assuming there is real interaction between methods and schools) will be more highly correlated with one another than with the other η -terms. We shall now assume that for our p methods (and the total population of schools) the η -terms are equally correlated, with correlation ρ ; we shall also assume that they are normally distributed with the same variance σ_{η}^2 for each method. The population mean will in each case be zero.

With this definition of our model the expected values of the mean squares for the analysis of variance are as in Table 1. For the benefit of the reader

TABLE 1

Variance	d. f.	Expected Value of Mean Square
Methods	$p - 1$	$\sigma_{\epsilon}^2 + n[\sigma_{\eta}^2(1 - \rho) + \sigma_{\xi}^2] + nq \sum_r A_r^2 / (p - 1)$
Schools	$q - 1$	$\sigma_{\epsilon}^2 + n[\sigma_{\eta}^2(1 - \rho) + \sigma_{\xi}^2 + p\rho\sigma_{\eta}^2] + np\sigma_{\beta}^2$
Methods \times Schools	$(p - 1)(q - 1)$	$\sigma_{\epsilon}^2 + n[\sigma_{\eta}^2(1 - \rho) + \sigma_{\xi}^2]$
Within Classes	$pq(n - 1)$	σ_{ϵ}^2

who is doubtful of the procedure for obtaining such a table, the derivation of the expected value of the mean square for methods is reproduced here.

The sum of squares between methods is given by

$$\sum_r qnM_r^2 - \frac{1}{pqn} \left(\sum_r qnM_r \right)^2 \quad (r = 1, \dots, p), \quad (6)$$

or more conveniently by

$$\frac{qn}{p} \sum_{k < l} (M_k - M_l)^2 \quad (k, l = 1, \dots, p), \quad (7)$$

where

$$\begin{aligned} M_k - M_l &= \frac{1}{qn} \sum_s \sum_t X_{kst} - \frac{1}{qn} \sum_s \sum_t X_{lst} \quad \left\{ \begin{array}{l} s = 1, \dots, q \\ t = 1, \dots, n \end{array} \right\} \\ &= (A_k - A_l) + \frac{1}{q} \sum_s [(\eta_{ks} - \eta_{ls}) + (\xi_{ks} - \xi_{ls})] \\ &\quad + \frac{1}{qn} \sum_s \sum_t (\epsilon_{kst} - \epsilon_{lst}). \end{aligned} \quad (8)$$

Substituting in (7), squaring out and taking the expected value of the resultant expression, we obtain

$$\frac{qn}{p} \sum_{k < l} \left\{ (A_k - A_l)^2 + \frac{2}{q} [\sigma_\eta^2(1 - \rho) + \sigma_\epsilon^2] + \frac{2}{qn} \sigma_\epsilon^2 \right\} \quad (k, l = 1, \dots, p),$$

which reduces to

$$qn \sum_r A_r^2 + n(p-1)[\sigma_\eta^2(1 - \rho) + \sigma_\epsilon^2] + (p-1)\sigma_\epsilon^2. \quad (9)$$

The required result follows.

If, instead of the above definition of β_s , we define β_s to be such that $(\mu + \beta_s)$ is the mean of the s th school over the p methods only, then $\sum_r \eta_{rs} = 0$ (as in the case of Model I). It is easy to show that ρ will now have the value $-1/(p-1)$. [If we substitute this value for ρ in Table 1, the expected value for the *schools* variance becomes $(\sigma_\epsilon^2 + n\sigma_\eta^2 + np\sigma_\beta^2)$, which does not contain σ_η^2 —a result which is obvious from the definition of β_s , now being assumed.] It might be argued that the correlation ρ is an artifact, since its value depends on the way the β -values are defined and ρ can thus be made to have almost any value we please. But the reader should note that correlations in the variance component analysis cannot be avoided when there is heterogeneity of correlation between methods [case (c) below].

We will now consider three possible sources of bias for the *methods v. interaction* F -test. [Bias arising from non-normality in the data will not be considered in the present paper. Much work has been done in this field and, while most of it has been concerned with the simpler applications of the analysis of variance and not with more complex analyses such as may occur in education, it is probably true to say that these findings have general application.] An application of Neyman's technique to the modified form of the basic model for each of the three cases is useless, since the B -ratio is found to be unity. The results are not reproduced here. It is possible, however, to make some fairly definite pronouncements on the bias involved in each case.

Case (a). Heterogeneity of variance within classes (from school to school)

As a result of an empirical study, Lindquist and Godard (12, pp. 139-144) concluded that this type of heterogeneity "will not seriously affect the validity of the test of significance of methods differences based on the ratio of the M and $M \times S$ variance." A corollary to this result is that heterogeneity of group errors from school to school will not seriously bias the F -test.

Case (b). Heterogeneity of variance within methods

This type of heterogeneity may arise in two ways: either (i) the variance within classes may vary from method to method; or (ii) the variance due to "real" interaction may vary from method to method.

It seems unlikely, as Lindquist remarks (12, p. 144), that the methods would produce sufficiently large differences in variability to disturb the F -test seriously. But, where this did happen, the following remarks about bias might be made:

(i) No bias results from this type of heterogeneity when only two methods are involved. This can easily be established analytically.

(ii) With more than two methods, the bias is likely to be positive. It is known that when a t -test is applied to two random groups of the same size, heterogeneity of variance causes the test to be positively biased (7, p. 170). There is no contradiction between this result and that stated in the previous paragraph. Heterogeneity of variance produces bias in the t -test when applied to random groups but not when applied to matched groups. The latter case corresponds to the replicated experiment with two methods.

It is very likely that the same holds for the F -test when applied to more than two heterogeneous groups; and, if so, it would also apply to the (M v. $M \times S$)-test (when more than two methods are involved). A consideration of special cases adds support to this conclusion.

A discussion of this type of bias for a similar situation in agricultural research is to be found in Cochran and Cox (4, pp. 396-398). A more general discussion of the problem is to be found in Cochran (2). As a possible method of dealing with heterogeneity of variance, Cochran suggests the separation of the methods comparisons into single comparisons and the computation of separate error terms for each. It is better, however, if such a solution is found to be unnecessary (involving as it does a considerable loss in degrees of freedom).

Case (c). Heterogeneity of correlation between class means (within methods)

The point to be noted here is that some methods may be more alike than others, and, consequently, their interaction effects (with schools) will be more closely related with one another than with those for the other methods—thus producing heterogeneity of correlation between the class means within methods.

The type of bias present can be easily demonstrated with fictitious data for a highly theoretical case. (The example which follows probably affords a better understanding of the way in which the bias operates than is to be gained by any lengthy analysis).

Consider an experiment involving two methods, A and B , and seven schools. Let the means of the experimental groups be as follows:

	Schools						
	1	2	3	4	5	6	7
Method A	39	55	45	47	46	40	51
Method B	38	40	46	40	38	43	42

Then, with equal numbers in the experimental groups, the *methods* and *interaction* components of the variance analysis read:

	d. f.	Sum of Squares	Variance	<i>F</i> -ratio
Methods	1	87.5	87.5	4.375
Methods \times Schools	6	120	20	

The *F*-ratio is not significant (for 1 and 6 d. f., $F = 5.99$ at the 5 per cent level of significance).

Now suppose that a third method, *C*, had been incorporated in the experiment and let us take the extreme case of *C* being identical with *B*. Also, let us imagine, to present the argument in its simplest form, that in this experiment there is no sampling error and that the interaction term consists only of real interaction. Then the means for the school groups subjected to method *C* will be the same as those for the groups undergoing method *B*. An analysis of variance for the three methods will therefore still give the same value for the *F*-ratio; but now with 2 and 12 d. f., significance is obtained at the 5 per cent level ($F = 3.88$).

We might consider what would happen with further replication of method *B*. Thus, with four replications, significance can be obtained at the 1 per cent level ($F = 4.22$ for 4 and 24 d. f.). Obviously, with the given form of analysis, the replication process increases the number of degrees of freedom without producing any real increase in the precision of the comparison of the methods. With a separation of the methods comparisons, such as Cochran suggests (see above), the spurious effect can be avoided.

The fact that methods are never identical and that sampling and other errors are always present does, of course, considerably reduce the amount of bias of this type which can occur. It is very probable that in most practical cases it is not serious. The use of covariance analysis or any other technique which improves precision by reducing random error will, of course, increase the importance of real interaction and so the type of bias under discussion. Covariance, etc., will also increase the effect of bias resulting from heterogeneity of variance of "real" interaction [see case (b) above].

Before concluding this section, two matters may be mentioned which are *not* irrelevant to the above discussion:

(i) As several writers have pointed out [e.g., Lindquist (12, p. 98); Webb and Lemmon (19)], similarities between methods may also operate in an *F*-test to mask other significant methods differences present, [i.e., speaking more technically, such similarities reduce the power of the *F*-test, cf. Johnson (9)]. Diamond (6) contends that the effect is normally small. It will be seen that, in the case of replicated methods experiment, both masking and case (c) bias may be present; it will also be seen that they are in opposition to

each other. Which predominates would depend on the relative importance of real interaction and random error.

(ii). It is to be observed that the analysis of variance of repeated measurements for a group of individuals is similar in form to that for the replicated methods experiment. *Individuals* correspond to *schools* and the sets of measurements (or *trials*) correspond to *methods*. Also the main F -test is *trials v. interaction* (individuals \times trials) corresponding to the (M v. $M \times S$)-test of the methods experiment.

It follows that a somewhat similar discussion of bias is involved. Case (a) does not arise but cases (b) and (c) are applicable.

It is very likely that the bias arising from heterogeneity of correlation between interaction effects can be more serious in the repeated measurements analysis than in the other. Since the sets of measurements must succeed each other in time, this is bound to result in greater correlation between sets of measurements coming close together than those further apart; the heterogeneity of correlation will increase the greater the intervals of time between the measurements. Lindquist (11) covers himself on this point when he states that his treatment of the analysis depends on the assumptions that all individual regression lines are linear and parallel and that deviations from individual regression are normally distributed and of equal variance for all subjects. He regards Alexander's tests (1) as superior in that they provide for the possibility of individual differences in regression. Certainly Alexander's method of analysis is able to reveal any heterogeneity of individual regression which may be present. But Lindquist does not make the obvious point that, as a result of this heterogeneity, Alexander can only apply his F -tests to study trend for the group he was considering and not for the larger population with which Lindquist was concerned. There would appear to be two alternatives: either (a) to apply Alexander's method and so make a study of trend for the group only; or (b) to apply the simpler method of Lindquist to obtain a generalized result with the knowledge that, in certain cases, the result may be seriously biased.

2. *Proportionate Numbers in Subclasses* ($n_{rs} = Na_{rs}$)

It is generally accepted that difficulties in the application of the analysis of variance arise only with disproportionate numbers in the subclasses (the *nonorthogonal* case) and that proportionate numbers involve no more than slight computational changes of the procedure for equal numbers per subclass. This point of view is quite legitimate in the case where the hypothesis being tested has reference only to the rows and columns—whatever they represent—involved in the experiment (Eisenhart's Model I falls into this category). But it is inaccurate in the type of experiment—common in education—where the object is a generalized result which applies to a larger population (i.e., where Eisenhart's Mixed Model or a similar model-type applies). Where the

interaction variance has other components of variance besides that due to the variance within subclasses, proportionate numbers in the subclasses will introduce bias into the F -test of *treatments* against *interaction*.

This type of bias would appear to have been discovered first by Smith (15), who gives the results of a variance component analysis of Eisenhart's Model II with proportionate numbers in the subclasses. Concerned as we are here with experimental designs common in education, it will be more instructive if we consider the results for the special mixed model which underlies the methods experiment replicated over a number of schools.

The model is

$$X_{rst} = \mu + A_r + \beta_s + \eta_{rs} + \xi_{rs} + \epsilon_{rst} \quad \left\{ \begin{array}{l} r = 1, \dots, p \\ s = 1, \dots, q \\ t = 1, \dots, n_{rs} \end{array} \right\}, \quad (10)$$

where the symbols have the same meaning as in subsection 1, but now, with proportionate numbers in the subclasses, n_{rs} can be written as $Na_r b_s$, where N is the total number of cases and the a 's and b 's represent the proportions of cases corresponding to columns (methods) and rows (schools), respectively, $(\sum_r a_r = 1 = \sum_s b_s)$.

Since we have already dealt with the problem of heterogeneity of variance and correlation in the previous subsection, we shall assume homogeneity of variance and correlation for the present version of our model. The results of a variance component analysis are then as shown in Table 2.

Applying the null hypothesis, namely,

$$A_k = A_l \quad (k, l = 1, \dots, p), \quad (11)$$

we obtain for the B -ratio of the $(M \text{ v. } M \times S)$ -test the expression

$$(q-1) \frac{N(1 - \sum_r a_r^2)(\sum_s b_s^2)S^2 + (p-1)\sigma_\epsilon^2}{N(1 - \sum_r a_r^2)(1 - \sum_s b_s^2)S^2 + (p-1)(q-1)\sigma_\epsilon^2}, \quad (12)$$

where

$$S^2 = [\sigma_\eta^2(1 - \rho) + \sigma_\xi^2]. \quad (13)$$

Subtracting the denominator from the numerator of this expression we obtain the quantity

$$\begin{aligned} & N(1 - \sum_r a_r^2)S^2[(q-1) \sum_s b_s^2 - (1 - \sum_s b_s^2)] \\ & = N(1 - \sum_r a_r^2)S^2[\sum_{u < v} (b_u - b_v)^2] \quad (u, v = 1, \dots, q), \end{aligned} \quad (14)$$

which is positive except for the case in which the b 's are all equal (when it becomes zero).

TABLE 2

Variance	d. f.	Expected Value of Mean Square
Methods	$p - 1$	$\sigma_e^2 + \frac{N}{p-1} \left(1 - \sum_r a_r^2 \right) \left(\sum_s b_s^2 \right) [\sigma_\eta^2 (1 - \rho) + \sigma_\xi^2]$ $+ \frac{N}{p-1} \sum_{k < l} a_k a_l (A_k - A_l)^2 \quad k, l = 1, \dots, p$
Schools	$q - 1$	$\sigma_e^2 + \frac{N}{q-1} \left(1 - \sum_s b_s^2 \right) \left\{ \left(\sum_r a_r^2 \right) [\sigma_\eta^2 (1 - \rho) + \sigma_\xi^2] + \rho \sigma_\eta^2 \right\}$ $+ \frac{N}{q-1} \left(1 - \sum_s b_s^2 \right) \sigma_\beta^2$
Methods \times Schools	$(p-1)(q-1)$	$\sigma_e^2 + \frac{N}{(p-1)(q-1)} \left(1 - \sum_r a_r^2 \right) \left(1 - \sum_s b_s^2 \right) [\sigma_\eta^2 (1 - \rho) + \sigma_\xi^2]$
Within Classes	$N - pq$	σ_e^2

An immediate conclusion to be drawn is that inequalities among the b proportions (i.e., the proportions for schools) introduce bias into the F -test. Also from the fact that the B -ratio is greater than unity, it is likely that the bias is positive (special cases confirm this).

Having discovered this bias, we must ask: how does it arise? It is obviously due to the fact that the use of unequal proportions of pupils results in unequal weighting of the η - and ξ -terms when, of course, they should receive the same weighting. Also, for the same reason, inequalities among the a 's must also produce bias although no indication of this is given by a consideration of the B -ratio alone. (It will be seen in fact that this effect is equivalent to that of heterogeneity of variance within methods).

How serious may the bias be? It will first be noted that, unlike the bias discussed in case (c) of subsection 1, the type of bias with which we are concerned here involves both the η - and ξ -terms which, together, are seldom negligible relative to the sampling error term (their relative effect will normally be increased by the use of covariance or a similar technique). Therefore, with large inequalities, the bias may be far from negligible.

An indication of the magnitude of the bias for unequal b proportions is the amount the B -ratio exceeds unity. This quantity can always be estimated for any practical case. To illustrate we shall use the data given by Lindquist in one of his examples (12, p. 120 *et seq.*).

$$N = 440 \quad p = 4 \quad q = 5$$

$$a_1 = a_2 = a_3 = a_4 = \frac{1}{4}$$

$$b_1 = \frac{10}{110}, \quad b_2 = \frac{30}{110}, \quad b_3 = \frac{19}{110}, \quad b_4 = \frac{35}{110}, \quad b_5 = \frac{13}{110}$$

The analysis of variance reads:

	d. f.	Sum of Squares	Variance
Methods	3	988.6	329.5
Schools	4	1748.3	437.1
Methods \times Schools	12	172.8	14.4
Within Classes	420	2981.5	7.1

The entries in the last column may be taken as estimates of the corresponding expressions in the last column of Table 1. Thus, by simple arithmetic we obtain the following estimates:

$$\sigma_e^2 = 7.1; \quad S^2 = [\sigma_\eta^2(1 - \rho) + \sigma_\xi^2] = .352;$$

$$\sigma_e^2 + \frac{N}{p-1} (1 - \sum_r a_r^2) (\sum_s b_s^2) S^2 = 16.6$$

$$B\text{-ratio} = \frac{16.6}{14.4} = 1.15.$$

Note that for the given values of the b 's, the value of the B -ratio cannot exceed 1.3, the value of $(q - 1) \sum b_s^2 / (1 - \sum b_s^2)$.

However, the deviation of the B -ratio from unity is not in itself a good measure of the magnitude of bias. Allowance must be made for the numbers of degrees of freedom involved in the F -test. The greater the numbers of degrees of freedom, the more important a given deviation becomes and vice versa.

It is now tentatively suggested that in all applications of the B -ratio technique, the magnitude of bias is best measured by the expression

$$\frac{|B - 1|}{F_{1\%} - F_{5\%}}, \quad (15)$$

where $F_{1\%}$ and $F_{5\%}$ represent the values of F at the 1 per cent and 5 per cent levels of significance for the given numbers of degrees of freedom. It might then be established empirically, for any given design or model, how great the value of this expression must be before the bias becomes serious.

For the type of model discussed in this subsection, the bias due to inequalities among the b proportions can to some extent be overcome by testing *methods* not against the *interaction* variance provided by the straight-forward analysis of variance but against the estimate of the expected value of the methods variance on the null hypothesis (i.e., for the given numerical example, against 16.6 instead of 14.4). [For a fuller account see Smith (15) and Cochran (3).]

3. Unequal (Disproportionate) Numbers in Subclasses

The literature on exact procedures for analyzing data of this type is now considerable. Tsao's paper (18) is probably the most rigorous and comprehensive. However, these methods involve much more computational labor than is demanded by the normal variance analysis. Also they have always been concerned with the testing of particular hypotheses, i.e., hypotheses having reference only to the rows and columns (whatever they represent) of the data to be analyzed (Eisenhart's Model I); and they have not, as yet, dealt with general hypotheses, i.e., hypotheses concerning a larger population of rows or columns (Eisenhart's Model II or Mixed Model). Investigations in educational research have therefore favored approximate methods of dealing with unequal numbers in the subclasses—at least where the criteria of applicability were satisfied. By far the most popular among these methods is Snedecor's Method of Expected Proportionate Frequencies (16, 17). In this section we will apply the B -ratio technique to investigate the bias which the use of this technique entails.

There will be the two cases to consider: (a) where the hypothesis tested applies only to the rows and columns of the data (Eisenhart's Model I);

(b) where a general hypothesis is tested, applying to a total population of rows or columns (Eisenhart's Mixed Model, etc.).

Case (a)

In this type of analysis, as was stated earlier in the paper, the two main F -tests are *interaction v. within subclasses*, and, when this test is not significant, *columns (or rows) v. within subclasses*. [Tsao (18) deals with other possible tests.] Before we proceed to derive the corresponding B -ratios, it is to be noted:

(i) Since Snedecor's method employs proportionate frequencies, the interaction term will not contain any component due to main effects (the characteristic of orthogonality), i.e., the interaction term is independent of the values of the main effects. Also the variance for columns will be independent of the main effects for rows and vice versa.

(ii) In deriving the first of the two B -ratios, we assume interaction to be zero; and in the case of the second, we not only make this assumption but we also assume zero differences between the main effects involved in the F -test.

It follows from (i) and (ii) that no serious loss of generality will be incurred (and a considerable saving in algebraic labor will be gained) if we straightway assume that interaction and the differences between main effects are zero; i.e., each observation, apart from a constant which we will here take to be zero, will consist only of sampling error and may be represented by

$$\epsilon_{rst} = \left\{ \begin{array}{l} r = 1, \dots, p \\ s = 1, \dots, q \\ t = 1, \dots, n_{rs} \end{array} \right\}, \quad (16)$$

where p denotes the number of columns, q denotes the number of rows, and n_{rs} denotes the number of observations in the subclass (r, s) .

It will be assumed that the ϵ 's in all subclasses may be regarded as random samples from an infinite population of ϵ 's of zero mean and variance σ_ϵ^2 (the usual assumption of homogeneity of variance). Thus, σ_ϵ^2 is the E. V. of the variance *within subclasses*.

Also let

$$\begin{aligned} Na_r &= \sum_s n_{rs} \\ Nb_s &= \sum_r n_{rs} \quad (r = 1, \dots, p; s = 1, \dots, q). \\ Nc_{rs} &= n_{rs} \end{aligned} \quad (17)$$

Now let us derive the E. V.'s of the different sums of squares contained in Snedecor's Method. The sum of squares *between subclasses* is given by

$$\sum_r \sum_s N a_r b_s \left(\frac{1}{n_{rs}} \sum_{t=1}^{n_{rs}} \epsilon_{rst} \right)^2 - N \left[\sum_r \sum_s a_r b_s \left(\frac{1}{n_{rs}} \sum_{t=1}^{n_{rs}} \epsilon_{rst} \right) \right]^2, \quad (18)$$

which has E. V.

$$\begin{aligned} \sum_r \sum_s N a_r b_s \frac{\sigma_\epsilon^2}{n_{rs}} - N \sum_r \sum_s a_r^2 b_s^2 \frac{\sigma_\epsilon^2}{n_{rs}} \\ = \sigma_\epsilon^2 \sum_r \sum_s \frac{1}{c_{rs}} (a_r b_s - a_r^2 b_s^2). \end{aligned} \quad (19)$$

The sum of squares *between columns* is given by

$$\sum_r N a_r \left(\sum_s \frac{b_s}{n_{rs}} \sum_{t=1}^{n_{rs}} \epsilon_{rst} \right)^2 - N \left[\sum_r \sum_s \frac{a_r b_s}{n_{rs}} \sum_{t=1}^{n_{rs}} \epsilon_{rst} \right]^2, \quad (20)$$

which has E. V.

$$\sum_r N a_r \sum_s b_s^2 \frac{\sigma_\epsilon^2}{n_{rs}} - N \sum_r \sum_s a_r^2 b_s^2 \frac{\sigma_\epsilon^2}{n_{rs}} = \sigma_\epsilon^2 \sum_r \sum_s \frac{a_r (1 - a_r) b_s^2}{c_{rs}}. \quad (21)$$

Similarly the E. V. of the sum of squares *between rows* is

$$\sigma_\epsilon^2 \sum_r \sum_s \frac{b_s (1 - b_s) a_r^2}{c_{rs}}. \quad (22)$$

By subtracting the sum of (21) and (22) from (19), we obtain the E. V. of the *interaction* sum of squares

$$\sigma_\epsilon^2 \sum_r \sum_s \frac{a_r b_s (1 - a_r) (1 - b_s)}{c_{rs}}. \quad (23)$$

It follows that the *B-ratio* for the *F-test interaction v. within subclasses* is

$$\frac{1}{(p-1)(q-1)} \sum_r \sum_s \frac{a_r b_s (1 - a_r) (1 - b_s)}{c_{rs}}. \quad (24)$$

What values will this expression normally have?

It will first be noted that, when $c_{rs} = a_r b_s$, the *B-ratio* is unity since

$$\sum_r \sum_s (1 - a_r) (1 - b_s) = (p-1)(q-1). \quad (25)$$

This is, of course, to be expected since we are then dealing with proportionate numbers in the subclasses.

When $c_{rs} \neq a_r b_s$, the *B-ratio* may be positive or negative but a limited empirical study would suggest that it is normally positive. This again is to be expected for two reasons: (i) On the average the expression $a_r b_s / c_{rs}$ is likely to be greater than unity, since for a given difference between c_{rs} and

$a_r b_s$, the expression will exceed unity by a greater amount when $c_{rs} < a_r b_s$, than it will be exceeded by unity when $c_{rs} > a_r b_s$. (ii) The B -ratio, in any particular case, may be regarded as a weighted mean of the pq values of the expression $a_r b_s / c_{rs}$ for that case; it, too, will tend to have a value greater than unity.

It is, therefore, suggested here that the use of Snedecor's Method will normally produce a positively-biased test of interaction. The amount of bias will be indicated by the deviation of the B -ratio from unity or, better, by the value of the expression suggested at the end of subsection 2.

The B -ratio for the columns *v. within subclasses* F -test is

$$\frac{1}{p-1} \sum_r \sum_s \frac{a_r b_s (1 - a_r) b_s}{c_{rs}}. \quad (26)$$

It will be apparent that exactly the same can be said about this ratio as for the other.

Before we finish with case (a), it may be of interest to examine Tsao's modification of Snedecor's Method (18). Tsao "questions the validity of retaining the within variance derived from the original data while the other variances are derived from the adjusted data." To judge from the simplified case with which he deals at the end of his article, he would adjust the sum of squares *within subclasses* to the value

$$\sum_r \sum_s N a_r b_s \sum_t \left(\epsilon_{rst} - \frac{1}{n_{rs}} \sum_t \epsilon_{rst} \right)^2, \quad (27)$$

which will have E. V.

$$\sum_r \sum_s N a_r b_s \frac{(n_{rs} - 1)}{n_{rs}} \sigma_e^2 = \left(N - \sum_r \sum_s \frac{a_r b_s}{c_{rs}} \right) \sigma_e^2. \quad (28)$$

That is, the E. V. of his adjusted variance within subclasses is

$$\frac{1}{N - pq} \left(N - \sum_r \sum_s \frac{a_r b_s}{c_{rs}} \right) \sigma_e^2 \quad (29)$$

and not σ_e^2 as for Snedecor's variance within subclasses.

If we are agreed that $a_r b_s / c_{rs}$ will on the average be greater than unity, it follows that the above E. V. will normally be less than σ_e^2 . It would appear therefore that Tsao's correction will on the average increase the bias of Snedecor's Method.

Case (b)

It will clarify the discussion if we think of the columns as methods and the rows as schools. Our problem then is to investigate the bias of the (M *v.* $M \times S$)-test when Snedecor's approximate method is applied.

Obviously a part of the bias produced will be of the type discussed in

subsection 2 (provided, of course, there is either real interaction or group error). The rest of the bias will be of the type discussed in case (a) above. In order to study the importance of the latter type of bias for the (M v. $M \times S$)-test, let us take the special case where there is no real interaction and error consists only of sampling error. Then, from the analysis in case (a), it will be seen that the B -ratio for the (M v. $M \times S$)-test is

$$\frac{1}{p-1} \sum_r \sum_s \frac{a_r b_s (1 - a_r) b_s}{c_{rs}} \bigg/ \frac{1}{(p-1)(q-1)} \sum_r \sum_s \frac{a_r b_s (1 - a_r)(1 - b_s)}{c_{rs}}. \quad (30)$$

Since both numerator and denominator may be regarded as weighted means of the same pq values of $a_r b_s / c_{rs}$, it follows that the B -ratio will vary about unity, the degree of variation diminishing with the increase in number of the rows and columns. Therefore, as far as this type of bias is concerned, it is likely that Snedecor's Method will generally provide a more valid F -test for case (b) than for case (a).

The complete B -ratio for Snedecor's (M v. $M \times S$)-test, when both types of bias are involved, can be written down without further calculation (cf. previous subsection). It is

$$(q-1) \frac{N(1 - \sum_r a_r^2)(\sum_s b_s^2)[\sigma_\eta^2(1 - \rho) + \sigma_\epsilon^2] + \sigma_\epsilon^2 \sum_r \sum_s \frac{a_r b_s^2}{c_{rs}}(1 - a_r)}{N(1 - \sum_r a_r^2)(1 - \sum_s b_s^2)[\sigma_\eta^2(1 - \rho) + \sigma_\epsilon^2] + \sigma_\epsilon^2 \sum_r \sum_s \frac{a_r b_s}{c_{rs}}(1 - a_r)(1 - b_s)}. \quad (31)$$

An estimate of the value of this expression can easily be found for a given case.

In examining the bias increased by Snedecor's Method, no mention has been made of the χ^2 criterion for the applicability of the method. Snedecor established this criterion by empirical methods. Obviously the B -ratio, or rather some such expression as (15), could be established empirically as an alternative criterion. In dealing with the type of analysis discussed under case (b), it is possible that this alternative might prove superior.

V. Summary of Results

The basic model is

$$X_{rst} = \mu + A_r + \beta_s + \eta_{rs} + \xi_{rs} + \epsilon_{rst} \quad \left\{ \begin{array}{l} r = 1, \dots, p \\ s = 1, \dots, q \\ t = 1, \dots, n_{rs} \end{array} \right\}.$$

For a complete description of this type of model see p. 233. The main F -test is that of *columns* (the main effects of which are represented by the A -terms) against *interaction* (columns \times rows).

1. *Equal Numbers in Subclasses* ($n_{rs} = n$)

Three possible sources of F -test bias were considered:

(a) *Heterogeneity of variance within subclasses (from row to row)*. There is no evidence of bias in this case. The same applies to heterogeneity of variance of the ξ -effects from row to row.

(b) *Heterogeneity of variance within columns*. No bias arises when only two columns are involved. When there are more than two columns, the bias is likely to be positive (for definition of positive and negative bias see p. 229).

(c) *Heterogeneity of correlation between β -effects (of one column with another)*. The bias in this case is positive. In the typical methods experiment replicated in a number of schools it is unlikely to be serious; but for an analysis of variance of repeated measurements the bias involved might be considerable.

2. *Proportionate Numbers in Subclasses* ($n_{rs} = Na_r b_s$)

For the given type of model (also for Eisenhart's Model II and Mixed Model), proportionate numbers in the subclasses produce bias, again of a positive character. The amount of bias depends on the degree of inequality among the a and b proportions; also on the magnitude of the η - and ξ -variances relative to the ϵ -variance. Gross inequalities in the proportions are obviously to be avoided in setting up experiments. A formula, of general application, is suggested for measuring the magnitude of bias.

3. *Disproportionate Numbers in Subclasses*

In this case, F -test bias was studied for Snedecor's Method of Expected Proportionate Frequencies. Eisenhart's Model I was considered as well as the mixed model stated above.

For Eisenhart's Model I, bias, if present, will normally be positive. When Tsao's modification of Snedecor's method is applied, it would appear that the bias will on the average be increased.

In the case of the Mixed Model, part of the bias arises in the same way as for Model I, but it is likely that, in general, it will not have the same importance. The other part of the bias is of the same nature as that discussed in section 2.

It is suggested that, for the Mixed Model, the expression for measuring bias proposed at the end of section 2, might prove superior to the χ^2 -criterion as a test of the applicability of Snedecor's Method.

REFERENCES

1. Alexander, H. W. The estimation of reliability when several trials are available. *Psychometrika*, 1947, 12, 79-99.
2. Cochran, W. G. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, 1947, 3, 22-38.
3. Cochran, W. G. Testing a linear relation among variances. *Biometrics*, 1951, 7, 17-32.

4. Cochran, W. G. and Cox, G. M. Experimental designs. New York: Wiley, 1950.
5. Crump, S. L. The present status of variance component analysis. *Biometrics*, 1951, 7, 1-16.
6. Diamond, S. Comment on "A qualification in the use of analysis of variance." *Psychol. Bull.*, 1952, 49, 151-154.
7. Edwards, A. L. Experimental design in psychological research. New York: Rinehart, 1950.
8. Eisenhart, C. The assumptions underlying analysis of variance. *Biometrics*, 1947, 3, 1-21.
9. Johnson, N. L. Alternative systems in the analysis of variance. *Biometrika*, 1948, 36, 80-87.
10. Kendall, M. G. The advanced theory of statistics, II. London: Griffin, 1946.
11. Lindquist, E. F. Goodness of fit of trend curves and significance of trend differences. *Psychometrika*, 1947, 12, 65-78.
12. Lindquist, E. F. Statistical analysis in educational research. New York: Houghton Mifflin, 1940.
13. McNemar, Q. On the use of Latin squares in psychology. *Psychol. Bull.*, 1951, 48, 398-401.
14. Neyman, J. Statistical problems in agricultural experimentation. *Suppl. J. roy. statist. Soc.*, 1935, 2, 107-154.
15. Smith, H. F. Analysis of variance with unequal but proportionate numbers of observations in the subclasses of a two-way classification. *Biometrics*, 1951, 7, 70-74.
16. Snedecor, G. W. The method of expected numbers for tables of multiple classification with disproportionate subclass numbers. *J. Amer. statist. Assoc.*, 1934, 29, 389-393.
17. Snedecor, G. W. Statistical methods. Ames, Iowa: Iowa State College Press, 1946.
18. Tsao, F. General solution of the analysis of variance and covariance in the case of unequal or disproportionate numbers of observations in the subclasses. *Psychometrika*, 1946, 11, 107-128.
19. Webb, W. B. and Lemmon, V. W. A qualification in the use of analysis of variance. *Psychol. Bull.*, 1950, 47, 130-136.
20. Welch, B. L. On the z-test in randomized blocks and Latin squares. *Biometrika*, 1937, 29, 21-52.

Manuscript received 6/18/53

Revised manuscript received 8/19/54

LEAST SQUARES ESTIMATES AND OPTIMAL CLASSIFICATION

HUBERT E. BROGDEN

PERSONNEL RESEARCH BRANCH
THE ADJUTANT GENERAL'S OFFICE
DEPARTMENT OF THE ARMY*

A simple algebraic development is given showing that criterion estimates derived by usual multiple regression procedures are optimal for personnel classification. It is also shown that, for any assignment of men to jobs, the sum of the multiple regression criterion estimates will equal the sum of the actual criterion scores.

In earlier papers (1, 2), the author contended that estimates of job proficiency derived by least squares estimates will place men in jobs in the most efficient way possible with the given predictor battery available, and that the average estimated job proficiency obtained by the use of such least squares estimates will equal the average actual job proficiency of assigned personnel. This paper will seek to establish these two points in a more rigorous fashion.

Definition of Symbols

C_{ij} = the performance of individual i in job j .

\hat{C}_{ij} = estimates of the C_{ij} , each derived by regression equations from the same battery of tests and the same universe of individuals. It is assumed that the zero- and higher-order regressions involving the tests and the C_{ij} are linear.†

\bar{C}_{ij} = the average C_{ij} value for a subset of individuals having the same pattern of scores on the battery of tests.

X = an allocation matrix with elements, x_{ij} , taking on values of zero and one. The x_{ij} entries for any individual have a single entry of one, and the x_{ij} entries for job j have Q_j entries of one. The remaining entries are zeros. The arrangement of ones in X corresponds to the placement of men in jobs. The use of X to symbolize any possible allocation of men to jobs is convenient and facilitates algebraic manipulation. In computing an allocation sum (to be defined), the cross-products of C_{ij} and x_{ij} are summed. When x_{ij} is one, the corresponding C_{ij} is included in the sum; when x_{ij}

*The opinions expressed are those of the author and are not to be construed as reflecting official Department of the Army policy.

†In practice, the C_{ij} would obviously not be available for all individuals in each job. Regression equations applying to the same universe can be estimated through a series of validation studies with a separate study being necessary for each job. In actual use the \hat{C}_{ij} could then be computed for each applicant in each job.

is zero the corresponding C_{ij} is excluded. Thus, X represents any arrangement of zeros and ones, good or poor, consistent with the limitations already imposed, except that such an arrangement must be based solely upon the scores on the battery of classification tests. In other words, X represents any allocation of men to jobs consistent with the conditions of the problem.

K_j = a set of constants, one for each job. The K_j 's are assumed to have numerical values such that, with allocation of each individual to the job in which $(\hat{C}_{ij} + K_j)$ is highest, the number allocated to each job will correspond to the number specified by the quota for that job.

X' = a particular X , with the x'_{ij} for each individual taking on a value of one for the job in which $(\hat{C}_{ij} + K_j)$ is highest. X' otherwise conforms to limitations imposed on X .

$\sum_{i,j} C_{ij}x_{ij}$ = the allocation sum. From the definition of an allocation matrix, it is evident that the allocation sum is equivalent to a simple sum, across all individuals, of the C_{ij} 's for the job to which each is assigned by a given allocation matrix.

Q_j = the quota for job j .

The Proof

We seek to demonstrate that

$$\sum_{i,j} \hat{C}_{ij}x'_{ij} = \sum_{i,j} C_{ij}x'_{ij} \geq \sum_{i,j} C_{ij}x_{ij}.$$

Consider a subset of individuals having an identical pattern of scores on the battery of tests basic to the \hat{C}_{ij} . Since we have specified that x_{ij} and x'_{ij} are to be based solely upon the test scores, it follows that both will remain constant in summing across individuals within such a subset. Then, for such a subset

$$\sum_{i,j} (\hat{C}_{ij} + K_j)x_{ij} = \sum_j (\sum_i \hat{C}_{ij}x_{ij} + \sum_i K_jx_{ij}) \quad (1)$$

$$= \sum_j (x_{ij} \sum_i \hat{C}_{ij} + \sum_i K_jx_{ij}). \quad (2)$$

Similarly, it follows that

$$\sum_{i,j} (\hat{C}_{ij} + K_j)x'_{ij} = \sum_j (x'_{ij} \sum_i \hat{C}_{ij} + \sum_i K_jx'_{ij}). \quad (3)$$

As N approaches infinity, the number in the subset approaches infinity. Now the criterion means of subgroups with identical score patterns are the basic data for graphic plotting of zero- and higher-order regression lines. If the regression system is linear, points representing the criterion means will fall on or near the regression lines. As the number in the subgroups approaches infinity the difference between \bar{C}_{ij} , the criterion mean for the subgroup, and \hat{C}_{ij} , the predictor value derived from a linear regression equation, will approach zero. Consequently, it is also true that, for the subset, $\sum_i C_{ij}$, the sum of the criterion scores, approaches equality to $\sum_i \hat{C}_{ij}$.

The basis for the equivalence of \hat{C}_{ij} and \bar{C}_{ij} within a subset having an identical pattern of scores might also be stated as follows: It is a basic principle of least squares prediction that the mean is the point at which the sum of the squares of the deviations is minimal. \bar{C}_{ij} , hence, is the best least squares estimate of the criterion scores of individuals with an identical pattern of test scores. If the regression system is linear, \hat{C}_{ij} also provides the best least squares estimate. Hence, as N approaches infinity, the two must coincide.

From our definition of X' , we know that, for such a subset

$$\sum_{i,j} (\hat{C}_{ij} + K_j)x'_{ij} \geq \sum_{i,j} (\hat{C}_{ij} + K_j)x_{ij}. \quad (4)$$

From equations 2 and 3,

$$\sum_i (x'_{ij} \sum_i \hat{C}_{ij} + \sum_i K_j x'_{ij}) \geq \sum_i (x_{ij} \sum_i \hat{C}_{ij} + \sum_i K_j x_{ij}). \quad (5)$$

Substituting $\sum_i C_{ij}$ for $\sum_i \hat{C}_{ij}$, we obtain

$$\sum_i (x'_{ij} \sum_i C_{ij} + \sum_i K_j x'_{ij}) \geq \sum_i (x_{ij} \sum_i C_{ij} + \sum_i K_j x_{ij}). \quad (6)$$

We may also write

$$\begin{aligned} \sum_j (\sum_i \hat{C}_{ij} x'_{ij} + \sum_i K_j x'_{ij}) &= \sum_j (\sum_i C_{ij} x'_{ij} + \sum_i K_j x'_{ij}) \\ &\geq \sum_j (\sum_i C_{ij} x_{ij} + \sum_i K_j x_{ij}). \end{aligned} \quad (7)$$

Since (7) holds for any subset, it holds in summing over all individuals.

In summing over individuals within any job, K_j is a constant and may be factored out. Both $\sum_i x'_{ij}$ and $\sum_i x_{ij}$ are, from the definition of X' and X , equal to Q_j . Hence, we have

$$\begin{aligned} \sum_j (\sum_i \hat{C}_{ij} x'_{ij} + K_j Q_j) &= \sum_j (\sum_i C_{ij} x'_{ij} + K_j Q_j) \\ &\geq \sum_j (\sum_i C_{ij} x_{ij} + K_j Q_j) \end{aligned} \quad (8)$$

or

$$\sum_{i,j} \hat{C}_{ij} x'_{ij} + \sum_j K_j Q_j = \sum_{i,j} C_{ij} x'_{ij} + \sum_j K_j Q_j \geq \sum_{i,j} C_{ij} x_{ij} + \sum_j K_j Q_j \quad (9)$$

and, consequently,

$$\sum_{i,j} \hat{C}_{ij} x'_{ij} = \sum_{i,j} C_{ij} x'_{ij} \geq \sum_{i,j} C_{ij} x_{ij}. \quad (10)$$

We have, then, established two generalizations. First, we have shown that, as N approaches infinity, the predicted criteria for a set of jobs derived by the use of linear multiple regression equations yields, upon assignment of men to jobs, an allocation sum that is equal to or higher than that obtained

by any other assignment of individuals to jobs that is based on the test scores. Second, we have shown that, for any given assignment of men to jobs, the allocation sum obtained when regression estimates of the criterion are used becomes, as N approaches infinity, identical with that obtained when the criterion scores themselves are used.

REFERENCES

1. Brogden, H. E. An approach to the problem of differential prediction. *Psychometrika*, 1946, 11, 139-154.
2. Brogden, H. E. Increased efficiency of selection resulting from replacement of a single predictor with several differential predictors. *Educ. psychol. Meas.*, 1951, 11, 173-195.

Manuscript received 9/29/54

Revised manuscript received 11/15/54

AN IMPROVED METHOD FOR TETRACHORIC r

W. L. JENKINS

LEHIGH UNIVERSITY

From the ratio of the cross-products of a fourfold table, with the application of two tabled corrections, tetrachoric r 's can be estimated with a mean discrepancy of less than .005 even when splits vary greatly from the medians. The necessary calculations can be handled by slide rule and the correction tables used without interpolation.

Davidoff and Goheen (1) have recently published a table for estimating tetrachoric r 's directly from the ratio of the cross-products of a fourfold table without correction. Unfortunately, the method gives accurate answers only when both distributions are split at approximately their medians. When the splits are not close to the medians, the obtained r 's are always biased in the positive direction. With some extreme splits, the positive bias amounts to .10, .15, or more.

However, it is possible to *correct* the obtained tetrachoric r 's by a method which is described and explained below.

Method and Example

1. Letter the fourfold table so that a is smaller than d and ad is greater than bc .

(c) 43	(d) 612
(a) 32	(b) 39

2. Compute the cross-products ratio ad/bc .

$$(32 \times 612)/(43 \times 39) = 11.68$$

From Table 1 find the uncorrected tetrachoric r for the nearest value of the cross-products ratio.

For 11.60, uncorrected $r = .756$.

3. Compute the two marginal splits $(a + b)/\text{total}$ and $(a + c)/\text{total}$.

$$\frac{32 + 30}{726} = .10; \quad \frac{32 + 43}{726} = .10.$$

TABLE 2

Base Correction

[illegible]

TABLE 3
Multipliers for Base Correction

	When larger split is .40 or less, use Difference between splits										When larger split is more than .40, use Smaller split																												
	00	05	10	15	20	25	30	35	40	45	50	Uncorrected r	When larger split is .40 or less, use Difference between splits										When larger split is more than .40, use Smaller split																
	00	05	10	15	20	25	30	35	40	45	50			00	05	10	15	20	25	30	35	40	45	50		00	05	10	15	20	25	30	35	40	45	50			
96	35	42	62	85	96	107	117	122	128	132	134	134	96	52	104	102	98	92	86	81	76	72	68	64	62	52	104	102	98	92	86	81	76	72	68	64	62		
94	41	48	66	87	97	107	116	121	127	131	133	133	94	50	104	101	96	90	84	79	73	68	66	64	61	58	50	104	101	96	90	84	79	73	68	66	64	61	58
92	47	54	70	88	98	107	116	120	125	130	132	132	92	48	103	99	94	88	82	77	71	65	62	60	58	54	48	103	99	94	88	82	77	71	65	62	60	58	54
90	53	60	73	89	99	107	115	119	124	128	130	130	90	46	101	97	92	86	80	74	68	62	59	56	54	51	46	101	97	92	86	80	74	68	62	59	56	54	51
88	60	66	77	90	100	107	114	118	122	127	128	128	88	44	99	95	89	83	77	72	65	59	56	53	51	48	44	99	95	89	83	77	72	65	59	56	53	51	48
86	66	71	81	92	101	107	113	117	120	125	127	127	86	42	97	92	87	81	75	69	62	57	53	50	48	44	42	97	92	87	81	75	69	62	57	53	50	48	44
84	71	76	84	93	101	106	112	115	118	123	124	124	84	40	94	89	83	79	72	66	60	54	50	47	45	41	40	94	89	83	79	72	66	60	54	50	47	45	41
82	76	81	87	94	102	106	110	114	116	121	122	122	82	38	91	87	81	76	70	64	57	51	47	43	42	38	38	91	87	81	76	70	64	57	51	47	43	42	38
80	81	85	90	95	102	106	109	112	114	118	119	119	80	36	89	84	78	73	67	61	55	48	44	40	39	35	36	89	84	78	73	67	61	55	48	44	40	39	35
78	86	89	92	96	102	105	107	110	112	115	116	116	78	34	86	81	76	71	65	58	52	46	41	38	36	32	34	86	81	76	71	65	58	52	46	41	38	36	32
76	90	93	94	97	102	104	105	108	109	112	113	113	76	32	83	78	73	68	62	55	49	43	39	35	33	29	32	83	78	73	68	62	55	49	43	39	35	33	29
74	94	96	96	98	102	103	104	105	106	108	109	109	74	30	80	75	70	65	59	53	46	40	36	32	30	26	30	80	75	70	65	59	53	46	40	36	32	30	26
72	97	98	98	99	101	101	102	103	103	104	105	105	72	28	76	72	67	62	56	51	44	37	33	30	27	23	28	76	72	67	62	56	51	44	37	33	30	27	23
70	100	100	100	100	100	100	100	100	100	100	100	100	70	26	72	68	64	59	53	48	41	35	31	27	24	20	26	72	68	64	59	53	48	41	35	31	27	24	20
68	101	101	101	100	99	98	97	97	97	97	96	96	68	24	68	64	60	56	50	45	38	32	28	25	21	17	24	68	64	60	56	50	45	38	32	28	25	21	17
66	102	102	102	100	98	96	95	94	93	93	92	92	66	22	64	60	57	53	47	42	35	29	25	22	19	15	22	64	60	57	53	47	42	35	29	25	22	19	15
64	103	103	102	100	97	94	93	91	90	90	88	87	64	20	60	57	54	49	44	39	33	27	24	20	16	12	20	60	57	54	49	44	39	33	27	24	20	16	12
62	104	104	102	99	95	92	90	88	86	86	84	83	62	18	55	52	50	45	40	35	28	24	21	18	14	10	18	55	52	50	45	40	35	28	24	21	18	14	10
60	104	104	102	99	93	90	88	85	83	82	80	78	60	16	50	47	45	41	36	32	25	21	19	16	12	8	16	50	47	45	41	36	32	25	21	19	16	12	8
58	104	104	101	97	92	88	85	82	79	78	76	74	58	14	45	42	40	37	32	28	22	19	16	13	10	6	14	45	42	40	37	32	28	22	19	16	13	10	6
56	105	104	100	95	90	86	82	79	75	74	72	70	56	12	40	38	35	33	28	24	19	17	14	11	8	4	12	40	38	35	33	28	24	19	17	14	11	8	4
54	105	103	99	94	88	84	79	75	71	71	68	66	54	10	35	33	30	28	24	20	15	13	12	9	6	2	10	35	33	30	28	24	20	15	13	12	9	6	2

the splits and the uncorrected tetrachoric r . (b) If the larger split is .40 or greater, find the multiplier at the intersection of the smaller split and the uncorrected tetrachoric r .

Since the larger split is less than .40 in this example, use the difference of zero and the uncorrected r of .756 to find the multiplier of .90.

5. Multiply the base correction by the multiplier to secure the final correction.

$$.103 \times .90 = .093$$

6. Subtract the final correction from the uncorrected r to secure the corrected tetrachoric r .

$$.756 - .093 = .663$$

Explanation

Tables 1, 2, and 3 are derived from Pearson's tables of normal correlation surfaces (2). For Table 1, cross-product ratios for median splits were computed for r 's of .05, .10, .15, \dots , .95, and a curve constructed relating r to the cross-product ratios. The figures given in Table 1 are scaled from this curve.

Securing Tables 2 and 3 required a number of replottings of the Pearson data. Pearson's tables are set up in 0.1σ steps; decimal steps of marginal proportions are needed. Accordingly, it was necessary to pick values that corresponded roughly to the desired marginal splits at various levels of r and obtain cross-product ratios. These were plotted and replotted until a family of curves was obtained that related the needed corrections to three variables: the two marginal splits and the uncorrected tetrachoric r .

Table 2 is scaled from the family of curves according to steps of the two marginal splits, but for a single value of uncorrected tetrachoric r (.70). Except for such inaccuracies as may be introduced through repeated replottings, these corrections are precise when the uncorrected tetrachoric r is .70.

To avoid having a book of such tables (one for each step of uncorrected tetrachoric r), it was necessary to resort to some approximations. When both splits are small (below .40) the correction depends chiefly on the difference between the splits and the uncorrected tetrachoric r . When either split is large (above .40), the size of the smaller split (rather than their difference) has the greater influence. Table 3 is set up accordingly, presenting multipliers to be applied to the base corrections of Table 2.

Empirical check

The adequacy of the method is shown by the results of an empirical check involving the recomputation of 500 r 's taken from the Pearson tables. Table 4 shows at the top the discrepancies of the uncorrected r 's (all positively biased) such as would be obtained if Table 1 were used without correction. At the bottom are shown the residual discrepancies after the corrections of Table 2 and Table 3 have been applied. Even without interpolation, 88 per cent of the residual discrepancies are less than .005. With interpolation this rises to 94 per cent.

TABLE 4
Empirical Check on the Adequacy of the Correction Method

Discrepancies BEFORE correction (all positive)									
True r	.000 to .020	.021 to .040	.041 to .060	.061 to .080	.081 to .100	.101 to .120	.121 to .140	.141 to .160	.161 and up
.10	30	15	1						
.20	23	19	11	3					
.30	16	16	11	5	4	2			
.40	11	12	11	10	6	3	2		1
.50	10	12	9	4	9	5	2	2	3
.60	9	11	7	6	7	4	4	3	5
.70	9	12	11	7	7	3	2	1	4
.80	9	12	6	6	3	4	4	3	
.85	12	8	7	6	6				
.90	12	9	3						

Discrepancies AFTER correction				
True r	Without interpolation		With interpolation	
	.005 or less	More than .005	.005 or less	More than .005
.10	55	1	56	0
.20	52	4	54	2
.30	49	5	52	2
.40	51	5	51	5
.50	49	7	52	4
.60	50	6	53	3
.70	48	8	54	2
.80	42	5	42	5
.85	28	11	34	5
.90	19	5	19	5

REFERENCES

1. Davidoff, M. D. and Goheen, H. W. A table for the rapid determination of the tetrachoric correlation coefficient. *Psychometrika*, 1953, 18, 115-121
2. Pearson, K. (Ed.) Tables for statisticians and biometricians, Part II, London Biometric Laboratory, Univ. College, 1931.

Manuscript received 2/26/54

Revised manuscript received 8/16/54

BOOK REVIEWS

BENJAMIN FRUCHTER. *Introduction to Factor Analysis*. New York: Van Nostrand, 1954. pp. xii + 280. \$5.00.

Several good books are already available in factor analysis. What claim can be made for another? Fruchter answers this in his preface. "These treatments have been found difficult by many otherwise competent students because of the mathematics and notation involved. It is hoped that this book will serve as an introduction to the subject and as a steppingstone to these more advanced texts."

The first four chapters provide a logical and mathematical introduction to factor analysis. Spearman's two-factor theory and its generalization to Holzinger's bi-factor method are discussed first. Cluster analysis is then considered as a means for understanding the logic of factor analysis. Next comes a chapter of "mathematics essential for factor analysis," including the basic matrix algebra operations and the geometry of rotation. This is followed by a chapter in which the basic equations of factor analysis are developed.

The next four chapters present the principal computational procedures. The diagonal and centroid methods are given in one chapter; the multiple-group and principal-axes methods in another; orthogonal rotation in a third; and oblique rotation in a fourth.

The final three chapters discuss (a) the interpretation of factors, (b) various applications of factor analysis, (c) some of the controversial issues in contemporary factor analysis. The book concludes with a useful bibliography of 700 titles covering principally the period from 1940 (the year of Dael Wolfe's review) to 1952.

Fruchter's statement of factor analysis differs in two main ways from the books already familiar to the readers of *Psychometrika*. First, his account is briefer and probably simpler than that of any of his predecessors; secondly, it has a better claim to be a textbook, less claim to be a personal statement.

Fruchter is undoubtedly right in saying that many otherwise competent students find factor analysis difficult because of the mathematics and notation. For many years to come, statements of factor analysis will be needed in which the approach is by means of the logic and calculations rather than by any rigorous mathematical development.

Fruchter stresses (a) the practical applications of factor analysis, and (b) the computations. Ten examples of the use of factor analysis are given in the chapter entitled "Applications in the Literature." These are of an interesting diversity, ranging from investigations of conditioned responses and rat maze learning to prepsychotic personality traits and Supreme Court voting records. *Q*- and *P*-technique are represented as well as *R*-technique. The chapter should be useful in reminding the psychological student that in studying factor analysis he must remain a psychologist. The computations in factor analysis are presented in detail in chapters 5 through 8. The various steps are itemized, and the instructions are for the most part clear and straightforward, so that the student who works diligently through the presentation should be able to calculate a factor analysis in a research of his own. The more experienced factor-analyst will probably be glad to have these step-by-step descriptions both for his own reference and for supplying to the student who seeks his aid.

A price is paid, naturally enough, for this emphasis upon learning by doing. For the most part the controversial issues of theory are eschewed. Key concepts are frequently introduced with so little discussion that the student may have trouble in seeing why the factor-analyst has adopted the particular procedure. For example, the account of communalities is brief and in my view very unsatisfying. The use of communalities is probably the factor-analytic procedure which has been most criticized by statisticians. The student whose knowledge is derived from this book will hardly be able to reply to any criticism.

The distinction between common and specific variance is initially made (p. 45) without any mathematical or logical reason being supplied for its adoption, and the brief discussions on pp. 46-47 and pp. 51-52 might well serve to confuse rather than clarify issues for the student. For one thing, Fruchter points out that communalities enable one to reproduce the correlations, and unities enable one to reproduce the original test scores; however, Fruchter provides no reason for preferring the former to the latter. For another, the information that specific variance is potentially common variance needs further development. As written at present, the distinction between the two types of variance is made to appear an entirely arbitrary one depending upon the particular selection of tests made by the investigator.

The discussion of orthogonal and oblique rotation is no more satisfactory. The distinction between simple axes and primary axes (i.e., factor structure and factor pattern) is deferred until the final chapter, which is a pot-pourri of theoretical issues set aside earlier. Yet it is doubtful whether the student will get any real understanding of the techniques of oblique rotation presented in an earlier chapter without knowledge of this distinction. Secondly, the controversy between those who favor orthogonal and those who favor oblique rotation is also held over to the final chapter. Even then the arguments for both sides are summarized very briefly, with Fruchter making no attempt to adjudicate upon the issues.

Let us next consider how this book differs from previous books. Each of these may have been referred to as a textbook, but invariably it has been a personal document as well. Thurstone's book, for instance, is primarily a statement of his original contributions and distinctive theories; little space is given to opposed views, except sometimes by way of rebuttal. Burt never allows his reader to forget that factor-analysts are by no means agreed in their theories and procedures and enters into logical and mathematical controversies with zest. Likewise, in Thompson, in Cattell, and in Holzinger and Harman space is found for personal contributions and points of view.

Perhaps "the battle of the schools" is ending in factor analysis. Fruchter's book has none of the intensity of debate characteristic of factor analysis in the thirties and the forties. Evidently many of the old disputes are settled. While the logic of factor analysis continues under discussion (as in Eysenck's and Hartley's recent articles), the degree of "reality" to be attributed to factors appears increasingly to be a metaphysical rather than a scientific issue.

For the already settled issues, Fruchter's avoidance of controversy is probably a strength. Factor analysis may have had overmuch of polemics in the past. It is in respect to the currently unsolved problems that Fruchter's approach seems to me a less happy one. The critical student who asks: "Is simple structure invariant?" or "Do the present tests of significance work?" or "How can we be sure that the rank of the matrix is reduced by the present means for estimating communalities?" does not get answers from Fruchter's text. Probably Fruchter cannot be expected to have answers to all of these, but at least they might have been indicated to be unsolved questions. The student who reads Fruchter alone can hardly know how many issues remain unsettled.

To summarize, Fruchter has set himself a limited objective. He has dealt very lightly with the mathematics and with the more theoretical issues of factor analysis. His emphasis is upon the calculations. Within these limits, Fruchter has done a good job. His survey is well balanced and impartial. For the student who needs to become familiar with the computations, the book will be very helpful. For the person who desires an understanding of factor analysis beyond that required for routine calculation, the book will not in itself be a sufficient guide.

ANNE ANASTASI. *Psychological Testing*. New York: Macmillan, xiii + 682, 1954. \$6.75.

The reviewer of a textbook serves essentially three functions. He attempts first to evaluate the soundness of the work from the point of view of accuracy, fundamental soundness, and good judgment in those areas where opinion rather than demonstrable knowledge is involved. Secondly, the reviewer must consider the book from the point of view of the audience for which it is intended and indicate whether he thinks it is suitable for the purpose stated. Finally, he must evaluate the book from the point of view of its original contribution to the total body of knowledge in the area covered.

Concerning the soundness of the book the reviewer finds remarkably little with which to take exception. The point of view presented is conservative and scholarly. With a few minor exceptions, the material seems to be accurate and precise. Where individual judgment and evaluation enter the picture, these judgments are on the whole conservative and, while pointing out weaknesses, tend for the most part to be favorable toward tests and test authors. While it is, perhaps, no truer in the field of testing than in other fields, it certainly can be said that the construction and publication of a test is an exercise in compromise between what is theoretically right and desirable on the one hand and what is practical and feasible in terms of time and expense on the other. While the author of this book is fully aware of the need for improvement and makes many suggestions as to how this may be brought about, there is nothing in the book to discourage the potential author from undertaking the construction of a new test or to discourage the test publisher from expanding his offerings.

The development of the book seems logical. Chapter II, "Principal Characteristics of Psychological Tests," lays the groundwork for what is to follow. Some psychologists might take exception to the definition of a psychological test as "essentially an objective and standardized measure of a sample of behavior" on the basis that this definition is too comprehensive, including as it does almost every possible variety of test. In the writer's opinion, achievement tests could be handled better as a separate category rather than as an aspect of psychological testing. Problems of reliability, validity, standardization, etc., are substantially different for achievement tests than for psychological tests in many instances. The American Psychological Association Test Standards Committee recognized this fact in leaving to the American Educational Research Association the production of a code for achievement tests.

Dr. Anastasi says that one can "consider all tests as behavior samples from which predictions regarding other behavior can be made. Different types of tests can then be characterized as variants of this basic pattern." She indicates further that one needs to be cautious in talking about measures of capacity, since capacity cannot be directly measured but can only be inferred from a measure of behavior. With this point of view, the writer of this review is in hearty agreement but he feels that the text has not gone far enough in indicating that many of the measures described are useful only if they are used to infer future behavior. This is certainly true of intelligence tests both of the general variety and the factor batteries and obviously true of prognostic and aptitude tests.

In the writer's opinion it may be considered one of the weaknesses of this text that insufficient attention is given to the basic problem of comparing such measures of capacity with subsequent measures of achievement. The problem of the criterion is discussed effectively but inadequate attention is paid to the problem of units in terms of which such pre- and post-measures can be compared. In fairness, it should be said that as much is done in this text as is generally done, perhaps more, in dealing with these problems. For example, a considerable section is devoted to expectancy charts, which is a noteworthy addition to what is ordinarily found in similar texts.

The sections of the book which deal with various types of tests are particularly

well done. The selection of tests used for illustrative purposes seems to be representative and sufficient information is given to provide the reader with a good notion of the various types of tests.

With regard to the evaluation of the book from the point of view of the *audience* for which it is intended, the writer cannot speak with such complete single-mindedness. Dr. Anastasi defines the audience as "the general student of psychology" and says further "Today, familiarity with tests is required not only by those who give or construct tests, but by the general psychologist as well." It is the considered judgment of this reviewer that this textbook cannot be read intelligently by psychology students taking a course in psychological testing without their having had at least an elementary course in statistics. Even with such a prerequisite, the book would appear to be more satisfactory for graduate rather than undergraduate classes and for students majoring in psychology rather than in education. This is contrary to the opinion stated by Dr. Anastasi in her preface where she says, "no previous knowledge of statistics is presupposed by the present text . . ." and "... for the benefit of students with no prior familiarity with statistics, however, all statistical concepts employed in the text have been explained and illustrated. Such statistical concepts have been introduced as they were needed and have been discussed within the appropriate context. Thus, they should appear more meaningful to the beginner than they would if segregated into a special 'statistical chapter.' " It appears to the writer that the section on reliability particularly and to some extent the sections on validity and norms will be completely incomprehensible to a person who has not had previous knowledge of basic statistics.

Dr. Anastasi indicates further that the book would be helpful to the practitioner in a number of fields, including the guidance counselor, school psychologist, psychometrist, personnel worker in business and industry and the clinical psychologist. With this point of view, the writer takes no exception. In fact, he would recommend the book as one which it would be very valuable for any practitioner to review and to have on his shelf for frequent reference purposes, especially if he has had a good grounding in statistics and elementary measurement.

As regards the third responsibility posed for the reviewer, namely, the evaluation of a book from the point of view of its original contribution, the writer of this review must conclude that there is little in this book that would appeal as being unique either in method, content, or emphasis. This can hardly be considered a serious indictment since originality is not the prime requisite of a good text. Original research, of course, is ordinarily reported in the professional journals or in professional papers, and in any generation the giants like Truman L. Kelley or Lewis M. Terman, whose books mark educational milestones, must necessarily be few in number.

*Test Service and Advisement Center
Dunbarton, New Hampshire*

Walter M. Durost



LOUIS LEON THURSTONE

Louis Leon Thurstone

With the death of L. L. Thurstone on September 29, 1955, psychology lost one of its greatest, a unique figure on the psychological scene and one to whom psychologists will always be indebted. If any psychologist of the past quarter century deserved to be called Mr. Psychological Measurement, it was he. His major professional objective coincided with that of the Psychometric Society and of *Psychometrika*, both of which were founded under his leadership: The development of psychology as a quantitative, rational science. By virtue of his own contributions and his influence on others, psychology has taken long steps in the direction of fulfillment of this objective. No major aspect of the field of measurement was untouched by him.

Louis Leon Thurstone was born May 29, 1887, in Chicago, where in later years he spent the greater portion of his professional life and achieved his greatest distinction, at the University of Chicago. His parents were of native, Swedish stock, his father's occupations being, in turn, military instructor, Lutheran pastor, editor, and publisher. Owing to a mobile family life, Thurstone went to school in Illinois; Mississippi; Stockholm, Sweden; and Jamestown, New York. He attended Cornell University, where he specialized in engineering. Considering the few instances of which the writer has known in which psychologists have started from a base of engineering training, he has often thought that we should be better off if more psychologists had taken that educational route.

It was during his engineering-school days that the problem of the learning curve, and hence psychology, caught Thurstone's attention. On graduation, however, he was offered a position in the laboratory of Thomas A. Edison, where he spent the year of 1912. During the next two academic years, he taught engineering courses at the University of Minnesota, and there began his study of experimental psychology. Graduate work followed at the University of Chicago. In 1915 he accepted an assistantship in the new and active laboratory established by Walter V. Bingham at the Carnegie Institute of Technology. He received his doctorate from Chicago, with a dissertation on the learning curve. His academic rise at Carnegie was something of a record. Beginning with the rank of instructor in 1917, with a promotion each year he became professor and head of the department by 1920.

The year of 1923-24 was spent in Washington, D. C., with the Institute for Government Research, an agency devoted to the improvement of civil-service practices. From that time on, Thurstone had considerable influence, directly or indirectly, upon civil-service procedures.

After his marriage in the summer of 1924 to Thelma Gwinn, Thurstone assumed his professorship at the University of Chicago. In the course of time,

he had much to do with initiating and setting the pattern for the University's distinguished Board of Examinations. In 1938 he was honored with the appointment as Charles F. Grey Distinguished Service Professor. In 1948 he was Visiting Professor at the University of Frankfurt, and in the spring semester of 1953 at the University of Stockholm. He retired from Chicago in 1952, at which time he became Research Professor and Director of the Psychometric Laboratory at the University of North Carolina, which was his professional affiliation at the time of his death.

His unquestioned creative productivity can perhaps be attributed to certain traits that seem to stand out—his dissatisfaction with the status of psychology as he found it, his keen analytical ability, and his independence and originality of thought. These qualities showed themselves in a number of ways. For example, his originality was demonstrated relatively early. While a college undergraduate he developed a novel method for trisecting an angle and published a paper on it in the *Scientific American*. By the time he graduated he had developed a motion-picture camera and projector of unusual design. It was this that brought him to the attention of Edison. His independence of thinking showed itself in the fact that he did not read widely in the psychological literature, as he was quite willing to admit, and in his general choice of some of the less popular subjects to which to devote his energies.

His dissatisfactions were many. He was discontented with the state of affairs he found in connection with psychological testing, where a rapidly growing practice seemed to have little or no underlying theory. It was in 1924 that he published his only attempt at general psychological theory, in his book entitled *The Nature of Intelligence*. More clearly, the need for test theory led him into the development of quantitative formulations, culminating in his multiple-factor analysis. His impatience with what he considered to be the severe limitations of classical psychophysics led to his development of the basic theory of psychological measurement as effected through human judgments. He regarded his law of comparative judgment as one of his most important achievements.

His performance as an analytical thinker was most clearly brought to the attention of the writer when, in the fall quarter of 1935, he was privileged to attend Thurstone's seminar. Whenever a new problem came up in the seminar, it was a revelation to observe him go to work on it. He very quickly went to the heart of the problem, singling out the important variables in a way that made the problem appear simple. Incidentally, attending the seminar then and later were a large number of students, some at the post-doctoral level, who have achieved distinction in the field of psychometrics in their own right.

Thurstone's many contributions are so well known to readers of *Psychometrika* that they need not be enumerated here. Although he has been held in great respect by those outside the field of measurement, many of his developments have not had the general impact that they should have had. Thurstone

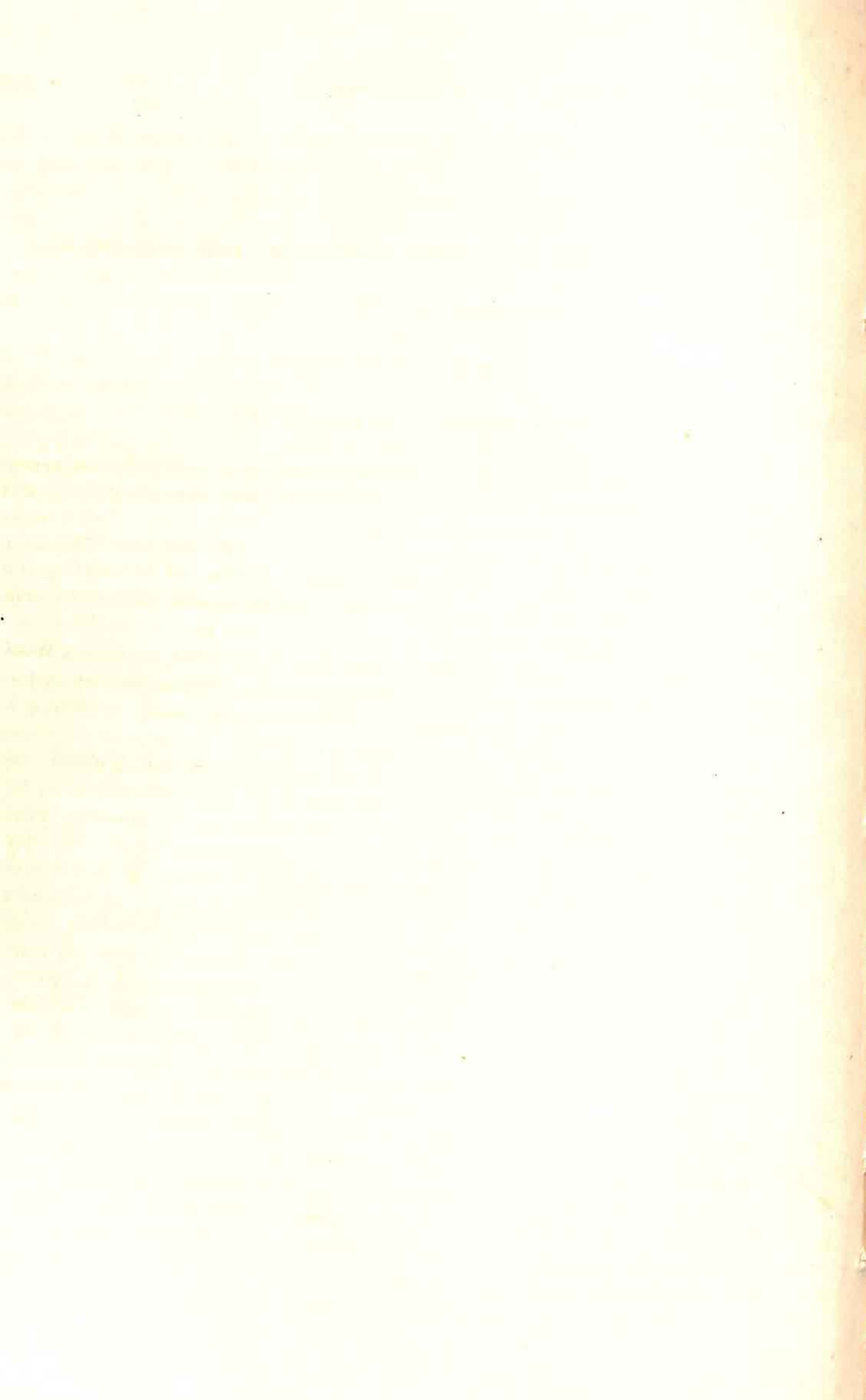
did not found a school, nor did he deal in popular or spectacular subjects. He often spoke to "deaf ears" because so few psychologists were prepared by virtue of interest or mathematical preparation to listen. Perhaps his contributions that have gained the widest notice are his attitude-scaling methods and attitude scales, and his discovery of primary mental abilities and his tests of them. Of the many quantitative methods that he has provided, probably that of multiple-factor analysis stands first by a wide margin in potential usefulness and impact upon psychology in general.

It is not so well known that in the later years of his life his energies were very much devoted to performance tests of personality. This interest sprang in part from dissatisfactions with both projective tests and personality inventories. He was also working on material for a book that was to give a general treatment of psychological measurement and on revisions of his tests of primary mental abilities. It is hoped that many of these last efforts had gone far enough to reach publication.

Thurstone was always quite willing to acknowledge the assistance from his wife, Thelma. They collaborated on many studies and for years were jointly responsible for the American Council on Education college-aptnitude examination. Their three sons are grown and started on their various professional careers. Besides the immediate family, Thurstone leaves behind a great many loyal and capable former students, who follow in the Thurstone-Chicago tradition, as well as many admirers around the world.

University of Southern California

J. P. Guilford



PSYCHOMETRIC THEORY: GENERAL AND SPECIFIC*

LEDYARD R TUCKER

PRINCETON UNIVERSITY
AND
EDUCATIONAL TESTING SERVICE

Fellow members of the Psychometric Society, and our colleagues, members of Division 5 of the American Psychological Association, I understand my mission tonight is to provide you with both humor and a mild message. Precedent has established an eminently high standard. Maybe I can explain the humbleness with which I approach this task by relating the account of an episode. This occurrence, like much illustrative data in *Psychometrika*, is fictitious, of course.

During my trip from the splendorous East Coast to the beauteous West Coast, I took vacation time to tour part of the magnificent country in between, as many of you, undoubtedly, did also. One afternoon I came upon a small tent village in a fairly remote mountain section. Not that such camps were so uncommon as to be remarkable in themselves; something about this particular camp caught my eye. In the center of this camp was a rustic conference table flanked on one side by a large blackboard mounted on cedar posts. This in itself was enough to raise a moderate interest, but my curiosity was intensely stimulated by another observation. Scribbled on this blackboard were a number of familiar symbols. Here was a sigma, there was a derivative, another place was a box labelled "factor matrix." How could this be? Needless to say I stopped to investigate. To my extreme pleasure I found myself among friends. A small group of Psychometricians had set up a summer seminar. There was nothing for me to do but to stay for dinner.

Before and during dinner we chatted about this and that. There were a number of requests for news such as: Were the Brooklyn Dodgers still running away with the race in the National League? Had Senator McCarthy started a new investigation? What were the leading plays in the summer theaters? After these matters had become settled, we sat back quietly for a period. This silence was ended by one member, whom, of course, I won't name, exclaiming: "Say, I've got one that should work. How about this? 751955." There was a short pause for ten seconds while the others seemed to stop and contemplate the proposition. Then there was loud and long laughter.

*Presidential Address to the Psychometric Society, September 5, 1955.

In the enthusiasm of the moment, another member called out "That would be like 683291." Again there was laughter. Later in these proceedings I was able to break in and ask for information. I was told that the group had found the off hours to become a bit boring after the first week of their conference. They had heard each other's jokes several times over and were tired of hiking and fishing. As an extra-curricular activity they had started to code the accumulation of humor. The numbers I had heard were such coded jokes. I decided to try this situation and so called out 10121492. There was not the least smile. Upon my inquiry of what was wrong, I wasn't given the hackneyed answer that "some could tell them and some could not." I had under-estimated our colleagues. They had not just gone about listing all jokes that they could think of and giving them a serial number in order of occurrence. They had gone about this activity like true Psychometricians. They had set up a system of analysis and classification of humor and had applied it to their sample of jokes. By use of systems of binary bits and addends of various other orders, they had established a system to encompass all possible jokes. Each sub-sub-etc. class, consisting of the jokes in one cell of this multiple classification system, had a unique identifying code which incorporated as meaningful units the position on each mode of classification. These identifications were then transformed to numbers in the decimal system. By considering the implications of any of the almost infinite number of empty cells they might conclude that the situation was humorous and have generated a new joke. This was what I had seen going on. It was a life-saver for the group, for otherwise they would not have been able to put up with each other for more than the second week. My weakness was that I had not realized the nature of their system and of its subtleties. I had pulled several boners. No wonder my selection was not humorous. Worst of all I had picked the lowest classification on the built-in criterion. I had told them that this joke was not the least bit funny.

With the field of humor now having been thoroughly studied, we may turn our attentions to other areas. For a short period I will direct my attention to material related to the title of this address, Psychometric Theory: General and Specific. Here I am taking the privilege of expressing my opinions on a recommended course of action for Psychometrics. My major concern is to maximize the extent to which observations of psychological phenomena can be validly matched with expectations obtained from rational theories.

While not wanting to dwell upon the philosophy of science—not that I would be able to do so if I wished—there are several aspects of my statement that warrant short discussion. By observations I mean the records obtained by definite procedures during the course of phenomena or of the observation of phenomena. The observations are subject to operational definitions. By expectations I mean statements describing the expected nature of observa-

tions. Such statements might vary on a continuum between vagueness and definiteness. I wish to consider definite statements as constituting definite expectations. Any vagueness could be considered as the introduction of approximation in the expectation. Thus, we might have approximate expectations.

I might illustrate by reference to color matching. The theory of the color pyramid would indicate that a given combination of color disks in particular proportions on one color wheel would be matched by another group of color disks on another wheel in given proportions. When the colors of these disks and the various proportions are specified, a definite expectation is produced. A corresponding observation could be obtained in the laboratory. A subject would be shown the color wheel with the standard combination of disks. The second wheel could be supplied with the other set of disks and the subject be asked to adjust the proportions until the color produced matched the color of the first wheel. Our observation would contain the specifications of the situation and a record of the proportions established by the subject.

I have avoided saying that observations were to be represented by rational theories. To make this latter statement would permit use of theories with infinite degrees of freedom so that these theories could be adjusted to represent any collection of observations, past, present and future. These theories would have no power, however, in providing definite statements of expectations concerning observations other than those observations the theories were adjusted to represent. In contrast, I have chosen to emphasize the correspondence between theoretically derived expectations and observations. The power of a theory is directly related to the number of definite expectations it produces and inversely to the errors between these expectations and the observations.

Let us consider the role of general psychometric theories in maximizing the extent to which derived expectations correspond to observations. Two aspects of this maximization appear: First, there may be a maximum correspondence, that is, minimum errors between the expectations and the observations. Second, the theory may yield a broad range of expectations. This is the place where general theories hold promise. We will all, undoubtedly, agree that it is desirable to have a systematic theory which will provide valid expectations for a large number of kinds of observations for many phenomena. There is a necessity, though, that we know how to use this theory to arrive at these expectations. Note my boner at the summer seminar previously described. As a further illustration, consider an advertisement that appeared a while ago in one of the weekly Princeton papers:

Lost: A black female cocker spaniel. Will not answer to "Daisy," although that is her name.

Unless each theory explicitly indicates phenomena and observations to which it is applicable, we may be in no better position than would someone searching for Daisy. I fear that in a number of instances we have mistaken vagueness for generality.

One may be interested in the relation between a method for analysis of data and a theory. In some senses a method of analysis might be considered as a general theory. A proper description of any such method includes a statement of the necessary conditions of situations in which the method might be employed. In a sense, however, we may contrast a method with a theory. This is in the specificity of definitions. If a construct might correspond to different kinds of observations in different situations, an ambiguity exists. A method that is applicable in several situations may be thought of as a theory in each situation. Each of these theories is separate from the other in that it includes the further definitions necessary to fit the method to the situation. Thus, a method may be considered as a family of theories. For example, multiple-factor analysis might be considered as a family of theories. Any time it is applied to a particular content area, it may be thought of as a theory. A method of analysis may be thought of as an abstraction from a number of situations with particular contents. In order to produce any particular theory from the family of theories represented by a method, it is necessary to specify particular contents.

We may be interested in classifying methods of analysis as more or less general. This classification may be taken to correspond to the number of different situations in which the method is applicable. A more general method would be applicable in a larger number of different situations. However, there may be a trend that the more general the method is, in this sense, the more extensive the definitions need be in order to make the method explicitly applicable to any one particular situation. Thus, the more general methods may be further from theories.

If a theory is to specify the observations and phenomena with which it deals, how then can a theory be general? We may wish to indicate the extensiveness of a theory's applicability by the contrasting terms: individual versus general. In case a theory concerns a large number of observables in a large variety of situations, the theory could be termed a relatively general theory. In case a theory concerns only a limited situation, the theory could be termed individual, or specific, to this situation. Note the particular usage of the word specific as synonymous to individual in this context. Both a general and an individual, or specific, theory should specify the observables with which it deals as well as situations or a situation in which the theory applies.

Although one might hope that a collection of individual theories could be amalgamated into a more general theory, there is some possibility that the individual theories may be too divergent to provide a basis for this

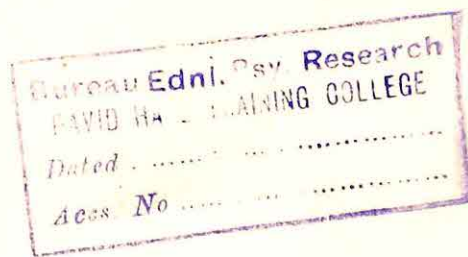
operation. Many individual theories arise from applications of psychometric methods to problems of applied psychology. The particular situations dealt with are defined by the needs of particular institutions. While this activity is a highly important aspect of psychometrics, its productivity of general theory is problematic. There is considerable danger that the practicing psychometrician may find himself in the kind of position illustrated by the following advertisement, which has also appeared in a weekly Princeton paper:

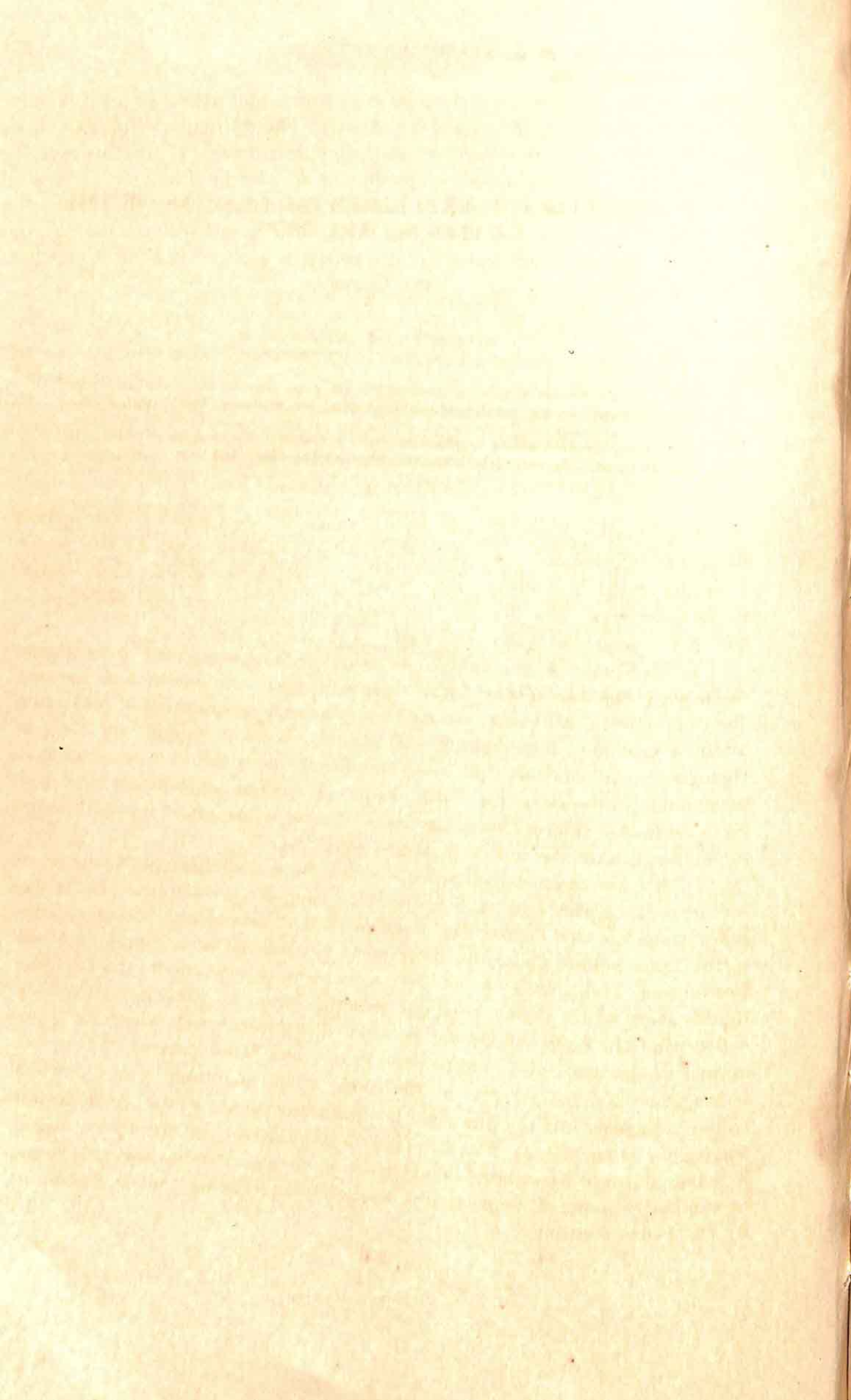
Aspiring Artist: Will decorate children's rooms. Funny animals a specialty, but will comply with any unreasonable request.

It seems to me we should be on guard against at least the extremes of such situations.

There seem to be problems, then, in our attempts at developing highly general theories. We may, in fact, produce instead of general theories families of individual theories which we might call psychometric methods. Or we may be so vague that the theory is unacceptable. Likewise, there are dangers in working with individual, or specific, theories. A recommended strategy is to attempt theories between these two extremes. The smaller, but not individual, theories may net us much in knowledge gained. These theories will be more easily established and experimented on by individuals. And finally we might hope that several such theories would be compatible for amalgamation with more general theories.

Manuscript received 8/10/55





F-TEST BIAS FOR EXPERIMENTAL DESIGNS OF THE LATIN SQUARE TYPE

NEIL GOURLAY

UNIVERSITY OF BIRMINGHAM

In an earlier paper, a method of analysis, due to Neyman and now known generally as variance component analysis, was used to examine *F*-test bias for experimental designs in education of the randomized block type. The same method is now applied to study *F*-test bias for designs of the Latin square type. The results, in general, disprove the view that, for a valid application of Latin square techniques, it is necessary that all interactions are zero.

In an earlier paper (5), a study was made of *F*-test bias for experimental designs in educational research of the randomized block type. In this paper, a similar study is made of those designs of which the simple Latin square is the prototype. The *B*-ratio technique, due to Neyman and described in the earlier paper, is again employed.

As McNemar (8) points out, the usual textbook statement of the theory underlying the use of the Latin square implies zero interaction between the main effects. McNemar claims that where the assumption of zero interaction is not met, investigators will obtain too many "significant *F*'s"; he then goes on to conclude that, since significant interactions are so common in psychological research, the Latin square is seldom appropriate and that "it is defensible only in those rare cases where one has sound a priori reasons for believing that the interactions are zero."

In the discussion which follows it will be shown that McNemar is by no means altogether correct in his point of view. It would appear that he has failed to realize that in the field of education and psychology, the application of the Latin square design has progressed beyond the usual simple textbook formulation and that some of the later applications show the need for modification of his rather sweeping generalization. In particular, it will be shown that the Latin square can be applied in several cases where the interactions are not zero; also, that in those cases where bias is present, it may well be negative and not positive as McNemar would maintain—a result which can only increase and not diminish the significance of any *F*-test. As defined in the earlier paper (5), an *F*-test is said to be positively or negatively biased, if, when the null hypothesis being tested is correct, it gives rise to a larger or smaller proportion, respectively, of significant *F*-ratios than is warranted by the *F*-distribution.

It will be helpful to the treatment of our problem if we distinguish between the two main types of interaction that can occur in psychology:

Type A—where each individual (or whatever the unit may be—class, grade, etc.) receives only one of the several treatments applied and is represented by only one measurement in the data to be analyzed. In this case, interaction is between main effects, such as treatments, schools, etc.

Type B—where repeated measurements are made on the same individual or group. In this case conditions may differ or different treatments may intervene; earlier measurements (treatments, etc.) may affect, i.e., interact with, those that follow.

The position is still further complicated in that *Type B* interaction may be accompanied by *Type A*. Also, besides pure interaction effects, the interaction component of any analysis of variance may contain other terms—described as group errors in the previous paper (5).

1. *Applications of the Latin Square Involving Type A Interaction Only*

1.1. *An experiment comparing several methods of teaching some school topic*

For simplicity let us take the case of a 3×3 square, say:

		Streams		
Schools	<i>f</i>	<i>u</i>	<i>v</i>	<i>w</i>
		<i>A</i>	<i>B</i>	<i>C</i>
	<i>g</i>	<i>B</i>	<i>C</i>	<i>A</i>
	<i>h</i>	<i>C</i>	<i>A</i>	<i>B</i>

i.e., in each school there are three experimental groups which are subjected to the three methods *A*, *B*, and *C* and which can be classified according to some other factor (e.g., streams). (For the benefit of some readers it is to be explained that in many English schools the children in each grade are assigned to classes according to level of ability. The process is known as *streaming*. A three-stream school is one in which there are three classes in each grade representing three levels of ability.) It will be assumed that the numbers in each group are equal—to avoid bias as discussed in the earlier paper (5).

Then we may consider two hypotheses: either (i) a particular hypothesis—the methods have the same mean effect when an average is taken for each method over the three schools and the three streams; or (ii) a general hypothesis—the methods have the same mean effect when an average is taken over the *total* population of schools (of which the three given schools are a random sample) and the three streams.

The first hypothesis is of little interest to the practical investigator. Furthermore, when real interaction is present between methods, schools and streams, the hypothesis cannot validly be tested by the Latin square design and analysis. There are obviously insufficient data: a factorial design is required. But it does not follow, as McNemar maintains, that the Latin square analysis would be positively biased. It might be positively biased in certain cases, but in others, it would be negatively biased, e.g., when real interaction exists only between two of the three classifications (the Latin square design then becomes effectively factorial).

Now let us consider the bias involved in the Latin square analysis when used to test the general hypothesis. It is now necessary to think of the three schools as a random sample from the *total* population of schools (which, as usual, will be taken as being infinite in number). Also the three schools should be assigned at random to the rows of the Latin square.

Then the mean scores for the nine experimental groups might be represented as follows:

$\pi_f + Q_u + T_A + \eta_{uA} + \epsilon_1$	$\pi_f + Q_v + T_B + \eta_{vB} + \epsilon_2$	$\pi_f + Q_w + T_C + \eta_{wC} + \epsilon_3$
$\pi_g + Q_u + T_B + \eta_{uB} + \epsilon_4$	$\pi_g + Q_v + T_C + \eta_{vC} + \epsilon_5$	$\pi_g + Q_w + T_A + \eta_{wA} + \epsilon_6$
$\pi_h + Q_u + T_C + \eta_{uC} + \epsilon_7$	$\pi_h + Q_v + T_A + \eta_{vA} + \epsilon_8$	$\pi_h + Q_w + T_B + \eta_{wB} + \epsilon_9$

(1)

where

(i) The general mean over the three methods, the three streams, and the total population of schools has been taken as zero.

(ii) The π -values are the main effects of schools, the π -value for each school being the mean for the school taken over the nine method-stream combinations. (It will be assumed that the total population of π -values has zero mean and variance σ_π^2 .)

(iii) The Q 's are the main effects of streams over the three methods and the total population of schools, and $\sum_3 Q = 0$, where $\sum_n F$ is defined to be the sum of all the terms or expressions of type F , and n denotes the total number of such terms.

(iv) The T 's are the main effects of the methods over the three streams and the total population of schools, and $\sum_3 T = 0$.

(v) The nine η 's represent the *real* interaction effects of the three methods and the three streams (over the total population of schools) and are such that

$$\begin{aligned} \sum \eta_{uk} &= 0 = \sum \eta_{rk} = \sum \eta_{wk} & (k = A, B, C), \\ \sum \eta_{lA} &= 0 = \sum \eta_{lB} = \sum \eta_{lC} & (l = u, v, w). \end{aligned} \quad (2)$$

(vi) The ϵ 's include *real* interaction between schools, on the one hand, and method-stream combinations, on the other, and group and sampling error. Each ϵ could in fact be expressed as

$$\epsilon = \zeta + \xi,$$

where ζ is the *real* interaction term for the school and method-stream combination to which the given ϵ belongs, and ξ is a purely random term made up of group error and sampling error.

The infinite population of ϵ 's (of ζ 's and ξ 's) corresponding to any one cell of the Latin square will have zero mean. We shall assume that the nine populations of ϵ -values (one for each cell) have the same variance σ_ϵ^2 (the usual assumption of homogeneity of variance).

The ϵ 's of the cells in any one row of the square will be correlated since their ζ -components are correlated. A few words of explanation might be helpful. For each school there are nine ζ -values (one for each method-stream combination) and their sum is zero; this follows from definition (ii) above. Also, for each method-stream combination there will be an infinite population of ζ -values (one ζ for each school). Furthermore, these nine populations of ζ -values will be correlated because the sum of the nine ζ -values for each school is zero. Since schools are assigned at random to the rows of the Latin square, it is not very difficult to see that, while the ζ -values which appear in the three cells of any one row will be correlated with one another, they will not be correlated with the ζ -values in the cells of the other two rows.

The correlations (nine in all) between the ϵ -values may not all be the same. This will happen when heterogeneity of correlation exists between the ζ -interaction effects. [A simple example of this type of heterogeneity is discussed more fully in the first paper (5).]

Let the correlations in the first row be denoted by ρ_{12} , ρ_{23} , ρ_{13} , where ρ_{12} represents the correlations between the ϵ 's in the first and second cells of the row and so on. Let the other six correlations be ρ_{45} , ρ_{56} , ρ_{46} and ρ_{78} , ρ_{89} , ρ_{79} . Also let $R = \sum_9 \rho$.

Now let us derive the E. V.'s (expected values) of the different sums of squares of the variance analysis over the *total* population of schools. The *total* sum of squares is

$$\sum_9 (\pi_f + Q_u + T_A + \eta_{uA} + \epsilon_i)^2 - \frac{1}{9} \left\{ \sum_9 (\pi_f + Q_u + T_A + \eta_{uA} + \epsilon_i) \right\}^2. \quad (3)$$

By equations above,

$$\sum_9 (\pi_f + Q_u + T_A + \eta_{uA} + \epsilon_i) = 3 \sum_3 \pi + \sum_9 \epsilon. \quad (4)$$

It follows that the E. V. of *total* sum of squares is

$$6\sigma_\pi^2 + 3 \sum_3 Q^2 + 3 \sum_3 T^2 + \sum_9 \eta^2 + 8\sigma_\epsilon^2 - \frac{2R}{9} \sigma_\epsilon^2. \quad (5)$$

The sum of squares *between methods* is

$$\sum_3 (M_A - M_B)^2, \quad (6)$$

where

$$M_A - M_B = (T_A - T_B) + \frac{1}{3}[(\epsilon_1 + \epsilon_6 + \epsilon_8) - (\epsilon_2 + \epsilon_4 + \epsilon_9)], \quad (7)$$

and

$$\text{E. V. } (M_A - M_B)^2 = (T_A - T_B)^2 + \frac{1}{9}[6\sigma_\epsilon^2 - 2\sigma_\epsilon^2(\rho_{12} + \rho_{46} + \rho_{89})]. \quad (8)$$

Therefore, the E. V. of sum of squares *between methods* is

$$\begin{aligned} \sum_3 (T_A - T_B)^2 + 2\sigma_\epsilon^2 - \frac{2R}{9}\sigma_\epsilon^2 \\ = 3 \sum_3 T^2 + 2\sigma_\epsilon^2 - \frac{2R}{9}\sigma_\epsilon^2 \quad (\text{since } \sum_3 T = 0). \end{aligned} \quad (9)$$

Similarly, the sum of squares *between streams* has E. V.

$$3 \sum_3 Q^2 + 2\sigma_\epsilon^2 - \frac{2R}{9}\sigma_\epsilon^2. \quad (10)$$

The sum of squares *between schools* is

$$\sum_3 (M_f - M_g)^2, \quad (11)$$

where

$$\begin{aligned} M_f - M_g = (\pi_f - \pi_g) + \frac{1}{3}[(\eta_{uA} + \eta_{vB} + \eta_{wC}) - (\eta_{uB} + \eta_{vC} + \eta_{wA})] \\ + \frac{1}{3}[(\epsilon_1 + \epsilon_2 + \epsilon_3) - (\epsilon_4 + \epsilon_5 + \epsilon_6)], \end{aligned} \quad (12)$$

and

$$\begin{aligned} \text{E. V. } (M_f - M_g)^2 = 2\sigma_\pi^2 + \frac{1}{9}[(\eta_{uA} + \eta_{vB} + \eta_{wC}) - (\eta_{uB} + \eta_{vC} + \eta_{wA})]^2 \\ + \frac{2}{3}\sigma_\epsilon^2 + \frac{2\sigma_\epsilon^2}{9}(\rho_{12} + \rho_{23} + \rho_{13} + \rho_{45} + \rho_{56} + \rho_{46}). \end{aligned} \quad (13)$$

It follows, on reduction, that the E. V. of the sum of squares *between schools* is

$$6\sigma_\pi^2 + \frac{1}{3} \sum_3 (\eta_{uA} + \eta_{vB} + \eta_{wC})^2 + 2\sigma_\epsilon^2 + \frac{4R}{9}\sigma_\epsilon^2. \quad (14)$$

The E. V. of the *residual* sum of squares can now be found by subtraction, but it can also be found directly after the manner of the other variance components. (The second procedure provides a check on the algebra in that the E. V. for *total* should equal the sum of the other E. V.'s). By either method the E. V. for *residual* is

$$\frac{1}{3} \sum_3 (\eta_{uA} + \eta_{vC} + \eta_{wB})^2 + 2\sigma_\epsilon^2 - \frac{2R}{9}\sigma_\epsilon^2. \quad (15)$$

It is to be noted that

$$\sum_9 \eta^2 = \frac{1}{3} \sum_3 (\eta_{uA} + \eta_{vB} + \eta_{wC})^2 + \frac{1}{3} \sum_3 (\eta_{uA} + \eta_{vC} + \eta_{wB})^2.$$

This is easily deduced from (2).

We are now in a position to apply the *B*-ratio technique to examine the bias involved in the *F*-test methods *v. residual*. The *B*-ratio for this test is obtained by (i) applying the null hypothesis that the main effects for methods are equal, i.e., $T_A = T_B = T_C (= 0$ since $\sum_3 T = 0$) and (ii) taking the ratio of the E. V. of the methods variance to the E. V. of the residual variance. It will be seen that it has the value

$$\left(1 - \frac{R}{9}\right) \sigma_e^2 / \left[\left(1 - \frac{R}{9}\right) \sigma_e^2 + \frac{1}{6} \sum_3 (\eta_{uA} + \eta_{vC} + \eta_{wB})^2 \right], \quad (16)$$

which will normally be less than unity if there is any real interaction between methods and streams. For the 3×3 Latin square, there are essentially two random arrangements:

$$\begin{array}{ccc} A & B & C \\ B & C & A \\ C & A & B \end{array} \quad \text{and} \quad \begin{array}{ccc} A & C & B \\ B & A & C \\ C & B & A. \end{array}$$

For the second arrangement, the *B*-ratio is of the form

$$\left(1 - \frac{R'}{9}\right) \sigma_e^2 / \left[\left(1 - \frac{R'}{9}\right) \sigma_e^2 + \frac{1}{6} \sum_3 (\eta_{uA} + \eta_{vB} + \eta_{wC})^2 \right].$$

It obviously leads to the same conclusions that apply to the first arrangement.

This suggests that the bias of the *F*-test will be negative and not positive as McNemar claimed. But another factor to be allowed for is the heterogeneity of correlation between the ϵ -values. As was shown in the previous paper, this heterogeneity produces positive bias. Whether the combined effect of the two types of bias is positive or negative will depend on various factors. But the point to be emphasised is that, for this application of the Latin square, the bias, if not negative, is at least less than that for the factorial type of experiment. If the bias is unimportant in the one case, it must *a fortiori* be unimportant in the other.

Unequal numbers of cases in the cells of the Latin square (whether proportionate or disproportionate with regard to the three classifications, schools, etc.) will also introduce bias into the *F*-test. This type of bias is discussed in the earlier paper for the case of the replicated (factorial) type of experiment.

1.2. *Alternative designs*

A single Latin square is not a practical design for a methods experiment. It provides too few degrees of freedom for the estimation of the residual or error variance. To obtain increased precision, two courses are available: either (1.2a) the replication of the same type of square; or (1.2b) the use of two or more different types of square with or without replication. In both cases, each square and each replication of it requires a separate sample of schools.

There would appear to be no special merit in preferring course (1.2a) to (1.2b) as some investigators have done—under the belief, presumably, that the use of as many different squares as possible is a necessary part of the randomization process. It is quite sufficient for randomization that the schools are allocated at random to the rows of a square [whether the same square throughout as in (1.2a) or to different squares as in (1.2b)].

1.2a. *The replication of the same square*

For simplicity, we will again take the 3×3 square discussed in section 1.1. Let there be n replications, involving, therefore, $3n$ schools in all. Then by an analysis very similar to that of section 1.1, it can be shown that the expected values for the different components of the variance analysis are as in Table 1 (same notation as before).

Two observations can be made: (i) By testing *methods* against *residual within schools* instead of *square residual*, a much more precise test is obtained. Furthermore, this F -test cannot be biased (negatively) as a result of any real interaction between methods and streams. It will, however, like the other test, be subject to any positive bias arising from heterogeneity of correlation between the cells in any one row. (ii) A test of interaction between methods and streams is provided by the F -tests: *square residual v. residual within schools* and *rows v. residual between schools*.

If interaction were shown to be present, and it were desirable that the methods should be compared for the three streams separately (instead of an average being taken as for the null hypothesis tested above), this could easily be achieved by analyzing the results for each stream separately. The precision of the tests involved would, however, be poor since school differences would be contained in the error variances.

1.2b. *Use of two or more types of square with or without replication*

An analysis similar in type to (1.2a) can again be carried out. The precise form the analysis takes will, of course, vary with the types of squares selected and the numbers of replications. To save space no such analysis is reproduced here. It might, however, be pointed out that, where 3×3 squares are involved, there are only two possible types of square and with the same number

TABLE 1

Variance	d.f.	Expected Value of Mean Square
Elements of Square	Methods	$(1 - R/9)\sigma_e^2$
	Streams	$(1 - R/9)\sigma_e^2$
	Rows	$(1 + 2R/9)\sigma_e^2 + (n/6) \sum_3 (\eta_{uA} + \eta_{vB} + \eta_{wC})^2 + 3\sigma_x^2$
	Square { Residual Uniqueness*	$(1 - R/9)\sigma_e^2 + (n/6) \sum_3 (\eta_{uA} + \eta_{vC} + \eta_{wB})^2$
Residual	Between Schools	$3(n - 1) (1 + 2R/9)\sigma_e^2$
	Within Schools	$6(n - 1) (1 - R/9)\sigma_e^2 + 3\sigma_x^2$

*The term *square uniqueness* was used by Corrigan and Brogden (3) in an analysis very similar to that above. Their type of application will be discussed later.

of replications of each, the analysis bears a certain similarity to that discussed on p. 283 *et seq.* (See also Table 2.) The conclusions to be drawn with regard to bias are the same as for case (1.2a).

1.3. More complicated applications

Consider the designs given in Figure 1.

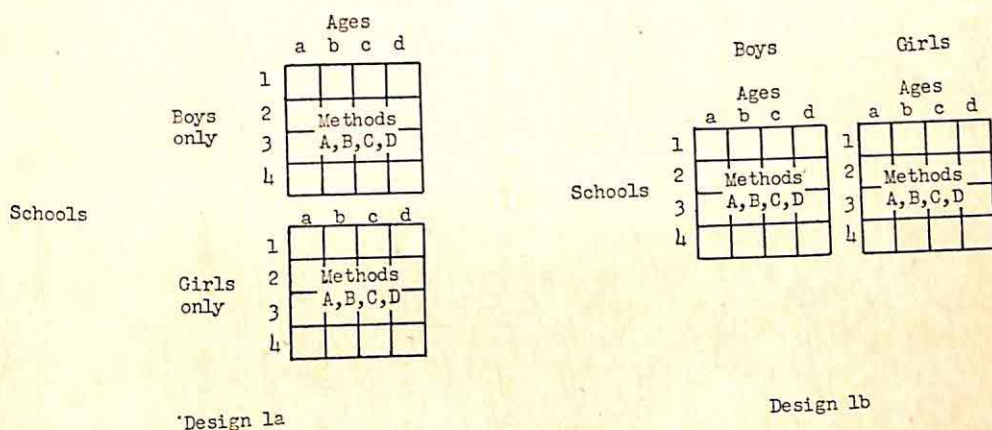


FIGURE 1

Design 1a, or something very similar, was used by Burt and Lewis (2). Once again it can be shown that the *B*-ratio for design 1a is less than unity, although as in all the other cases discussed, positive bias can result from heterogeneity of correlation between the four cells in each of the eight rows.

But the same cannot be said for design 1b. In this case it can be shown that with real interaction present, the B -ratio is likely to be greater than unity: a more serious degree of positive bias may, therefore, be present in the F -test.

2. Applications of the Latin Square Involving Type B Interaction

Most of the Latin square studies reported in journals have been of the type where the subjects of the experiments have been subjected to a succession of treatments and tested for each treatment: they could therefore involve Type *B* interaction [cf. Thomson (11), Sutherland (10), Grant (6), Edwards (4), and Archer (1)].

The Latin square design requires the order of succession of treatments to vary for the individuals and thus makes the problem of interaction more complicated than that which deals with Type *A* interaction only.

2.1. The case of a single Latin square

For example, consider three individuals subjected successively to three treatments *A*, *B*, and *C* in the following orders:

	1	<i>A</i>	<i>B</i>	<i>C</i>
Individuals	2	<i>B</i>	<i>C</i>	<i>A</i>
	3	<i>C</i>	<i>A</i>	<i>B</i>

With interaction present (whether type *A* or *B*), as stated earlier, the square does not provide sufficient data to test any worth-while hypothesis for the three treatments and the three individuals.

Can a general hypothesis be tested? Is there any difference between treatments when averaged over the total population of individuals of which our three cases are a sample? Once again the data are insufficient if type *B* interaction is present. The square involves only three of the possible six orders for the treatments, and all six orders would have to be considered in order to obtain a worth-while generalized result. We will examine the latter possibility presently. The only conclusion to be drawn is that a single Latin square is of little use when type *B* interaction is present or suspected. This, of course, is in accord with McNemar's point of view.

2.2. Replication of the same square

In this case each of the treatment sequences is applied not just to one individual but to a group of individuals. The applications of both Thomson (11) and Sutherland (10) fall into this category. (Sutherland's square is a Greco-Latin square but the same principles apply.) The method was also used by Corrigan and Brogden (3), whose application was discussed by Grant (6). Edwards (4) also illustrates the method.

In outward form, the analysis is very similar to that already discussed on p. 279. *Schools* is replaced by *individuals* and *rows* now represents groups of individuals undergoing different treatment sequences. *Streams* is replaced by some other classification. But there are some important differences which, to save space, we will only indicate:

(i) Besides possible type *A* interaction terms, there may be also type *B* interaction terms affecting each of the first four variance components (see Table 1).

(ii) In the earlier analysis, group errors (i.e., errors, other than random error, peculiar to a school group) were included in σ_e^2 and, therefore, affected all variance components. But in the present case, it will be seen that group errors will vary from cell to cell of the Latin square but will be the same for all the individual measurements in a cell, i.e., group error variance will form part of the first four variance components of the analysis but not of the two residual variances.

What tests may be applied and what is the position with regard to bias?

(i) An important test is that of *square residual* (or *uniqueness*) against *residual within individuals*. If significant, this may indicate either type *A* or *B* interaction, or group errors, or some combination of the three; the analysis cannot differentiate. With a significant result, there is little point in proceeding further. The test *treatments v. residual within individuals* would have such a limited interpretation that it would be virtually valueless; as a test of a general hypothesis about treatments, the test *treatments v. square residual* would be biased to an unknown extent.

(ii) If non-significance is obtained for the test of *square uniqueness*, a further test of zero interaction (and group error) is provided by taking *rows* (*groups* or *sequences*) against *residual between individuals*, provided the groups were random in the first place.

(iii) With both these tests non-significant, the other main effects might be tested against *residual within individuals*. But the reader must be warned against following such a test sequence blindly. It is to be remembered that a statistical test cannot *prove* the null hypothesis on which it is based; although the two preliminary tests are non-significant, it may still be the case that interaction (and/or group error) is present. A priori knowledge as to the likelihood of interaction and group error is obviously important. Where past experience would suggest that no interaction or group error is likely to be present, and the two preliminary tests confirm this, the tests of main effects against *residual within individuals* can be made with some safety. But where interaction or group error is known to be likely, little reliance can be placed on the tests of main effects, even though the preliminary tests give a non-significant result. In other words, the given experimental design is almost useless for dealing with this situation.

Corrigan and Brogden's data (3) show non-significance for both preliminary tests. Sutherland's data (10) show significance for the second test (the groups were not random) and would appear to show significance also for the first test; this would, of course, invalidate his other tests.

Edwards (4, p. 325) seems to regard *square residual* and *residual within individuals* as estimates of the same variance but, as we have seen, this can only be the case when interaction and group errors are zero.

2.3. Analysis involving complete sets of squares

When type *B* interaction is present or suspected, it is obvious that all possible treatment sequences must be considered if a generalized result is to be obtained. Grant (6) discusses this case.

Once again, for simplicity, we will consider the case of the 3×3 square. We will assume that individuals are assigned at random to the rows. There are then effectively only two Latin squares involved, corresponding to the six possible orders of treatment *ABC*, *BCA*, *CAB* and *ACB*, *BAC*, *CBA*.

For convenience we will give the eighteen cells of the design the numbers 1 through 18, in the order just stated, for the six sequences.

We will consider the case of n replications of the design, i.e., n individuals will be assigned to each row (or sequence). Then each of the n entries in any one of the eighteen cells may be represented as the sum of six terms of the form

$$\pi + Q + T + \eta + \xi + \epsilon, \quad (17)$$

where

(i) The general mean over the eighteen cells and the total population of individuals (assumed infinite) has been taken as zero.

(ii) The π -values ($6n$ in all) are the main effects of individuals averaged over the six sequences. (It will be assumed that the total population of π -values has zero mean and variance σ_π^2 .)

(iii) The Q 's (Q_u, Q_v, Q_w) are the main effects of columns, averaged over the six sequences and the total population of individuals, and $\sum_3 Q = 0$.

(iv) The T 's (T_A, T_B, T_C) are the main effects of treatments averaged over the six sequences and the total population of individuals, and $\sum_3 T = 0$.

(v) The η 's (18 in all) represent the joint effect of type A and B interaction for the 18 cells and are such that the six sums of six η -terms corresponding to the three treatments and the three columns are each zero.

(vi) The ξ 's (18 in all) represent possible group or cell errors. (It will be assumed that they are random and that the total population of ξ -values has zero mean and variance σ_ξ^2 .)

(vii) The ϵ 's ($18n$ in all) represent the residuals within cells after all the other effects have been taken out. It will be assumed that the total population of ϵ 's for each cell has zero mean and variance σ_ϵ^2 . Then the ϵ 's of the cells in any one row will be correlated. Let $\rho_{12}, \rho_{23}, \rho_{13}$ represent the correlations for the first row and so on. Also let $R = \sum_{18} \rho$.

The derivation of the E. V.'s of the different sums of squares in the variance analysis is not reproduced here but the results are given in Table 2.

What conclusions are to be drawn from this analysis?

(i) If one or both of the F -tests *residual between cells v. residual within individuals* and *rows (or sequences) v. residual between individuals* is significant, non-zero interaction and/or group errors is indicated; the analysis cannot differentiate. It would then be invalid to test *treatments (or columns)* against *residual within individuals* unless there was other evidence (possibly arising from the design of the experiment) to show that group errors were not present.

If, on the other hand, the first two tests were non-significant, the test of the significance of *treatments (or columns)* might safely be made. (The same warning as appears on p. 283 applies here).

(ii) It is always possible to test *treatments (or columns)* against *residual between cells*. The B -ratio for this test is never greater than unity. In this

TABLE 2

Variance	d.f.	Expected Value of Mean Square
Cells	Methods 2	$(1 - R/18)\sigma_e^2 + n\sigma_\xi^2 + 3n \sum_3 T^2$
	Columns 2	$(1 - R/18)\sigma_e^2 + n\sigma_\xi^2 + 3n \sum_3 Q^2$
	Rows 5	$(1 + R/9)\sigma_e^2 + n\sigma_\xi^2 + (n/15) \sum_6 (\eta_1 + \eta_2 + \eta_3)^2 + 3\sigma_\pi^2$
	Resid. between cells 8	$(1 - R/18)\sigma_e^2 + n\sigma_\xi^2 + (n/8) \left[\sum_{18} \eta^2 - \left(\frac{1}{3}\right) \sum_6 (\eta_1 + \eta_2 + \eta_3)^2 \right] + 3\sigma_\pi^2$
Residual	Between individuals $6(n - 1)$	$(1 + R/9)\sigma_e^2$
	Within individuals $12(n - 1)$	$(1 - R/18)\sigma_e^2$

respect the present analysis differs from that for the replication of the same square (see previous section), where type *B* interaction may affect both *treatments* (or *columns*) and *residual between cells* to give a *B*-ratio (and therefore a bias) of unknown size.

Before concluding this section it might be of interest to mention that type *B* interaction is similar to the carry-over effect studied by statisticians in animal science [cf. Patterson (9) and Lucas (7)]. Further, the experimental model which they consider is very much the same as that treated in this section. Their analysis of variance, however, follows quite a different pattern and permits the testing of a wider range of hypotheses. One of their findings is that there is no bias involved in the testing of *unadjusted* direct effects against the error variance. This agrees with conclusion (i) above. (It must be noted that group error does not occur in the animal science experiment.)

2.4. *More complex designs.*

No attempt will be made to consider *F*-test bias for analyses of more complex designs. It should now be apparent that where tests rest on the assumption of zero interaction and group error, the design should provide a test of this assumption. Also, in cases where such a test proves significant (or where the presence of interaction or group error is known to be likely even though unrevealed by any test), the design should furnish tests of main effects, which, although less precise than those which might otherwise have been used, possess *B*-ratios not exceeding unity.

Archer (1) shows himself to be aware of the limitations of the designs he offers in his paper, but considers that the difficulty could be partially overcome by ensuring that the interactions, for which his methods provide no test, are those which the investigator has decided a priori to be unimportant. There is a danger that these a priori decisions may be purely ad hoc assumptions and bear little relation to actual fact.

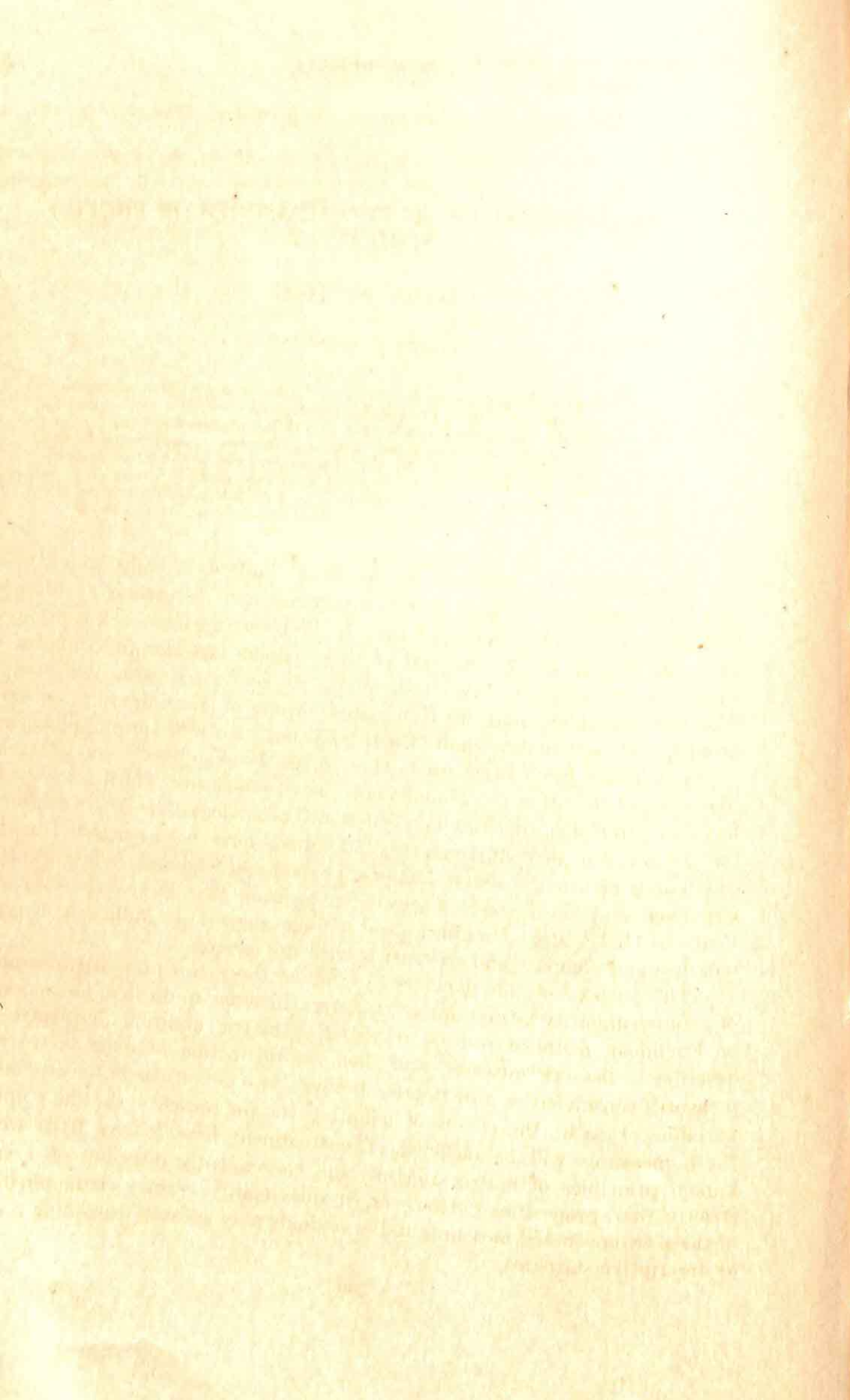
REFERENCES

1. Archer, E. J. Some Greco-Latin analysis of variance designs for learning studies. *Psychol. Bull.*, 1952, 49, 521-537.
2. Burt, C. and Lewis, B. Teaching backward readers. *Brit. J. educ. Psychol.*, 1946, 16, 116-132.
3. Corrigan, R. E. and Brogden, W. J. The trigonometric relationship of precision and angle of linear pursuit-movements. *Amer. J. Psychol.*, 1949, 62, 90-98.
4. Edwards, A. L. Experimental design in psychological research. New York: Rinehart, 1950.
5. Gourelay, N. *F*-test bias for experimental designs in educational research. *Psychometrika*, 1955, 20, 227-248.
6. Grant, D. A. The Latin square principle in the design and analysis of psychological experiments. *Psychol. Bull.*, 1948, 45, 427-442.
7. Lucas, M. L. Bias in estimation of error in change-over trials with dairy cattle. *J. agr. Sci.*, 1951, 41, 146-148.

8. McNemar, Q. On the use of Latin squares in psychology. *Psychol. Bull.*, 1951, 48, 398-401.
9. Patterson, M. D. The analysis of change-over trials. *J. agr. Sci.*, 1950, 40, 375-380.
10. Sutherland, J. An investigation into the prognostic value of certain arithmetic tests at the age of eleven plus. *Brit. J. Psychol., Statist. Sect.*, 1952, 5, 189-196.
11. Thomson, G. H. The use of the Latin square in designing educational experiments. *Brit. J. educ. Psychol.*, 1941, 11, 135-137.

Manuscript received 6/18/53

Revised manuscript received 8/19/54



CHARACTERISTICS OF TWO MEASURES OF PROFILE SIMILARITY

CHESTER W. HARRIS

UNIVERSITY OF WISCONSIN

Analogues of Pearson's coefficient of racial likeness and of Mahalanobis' distance measure have been proposed as descriptive statistics for comparing two individuals. This paper shows that two different definitions of "uncorrelated" variables—one associated with an inverse transformation and the other with a principal-axis transformation—give rise to these two descriptive statistics. The effects of putting the data into certain forms, such as equalizing the variances of the variables or equalizing the means of the persons, prior to using either of the two transformations, are discussed.

The interest in measures for assessing similarity (or dissimilarity) of profiles is reflected in such recent summaries and discussions as those of Osgood and Suci (5), Gaier and Lee (3), Webster (8), Cronbach and Gleser (2), and Thorndike (7). Several of these papers consider the problem of similarity of profile for two individuals, as contrasted with two groups. The latter problem may be formulated as one of discrimination between groups, and, as Cronbach and Gleser point out, two well-known approaches to its solution have been tried. One is the Pearson coefficient of racial likeness and the other the Mahalanobis distance measure, which is known to be related to Fisher's discriminant function. The analogues of these two measures for the problem of comparing two individuals have been suggested in the discussions mentioned above. For one, Osgood and Suci and, independently, Cronbach and Gleser have suggested a measure that is analogous to the Pearson CRL; also, Cronbach and Gleser suggest a Mahalanobis-type measure and compare and contrast it with the former.

The purpose of this paper is to examine these two proposed measures of profile similarity as descriptive statistics. In order to do this, the concept of Euclidean distance will be reviewed, a matrix notation developed to describe a distance measure, and then the distinction between these two measures considered as a distinction between two definitions of uncorrelated variables. Finally, the effects of adopting certain forms of the data upon these measures will be outlined. The treatment here follows from well-known principles of matrix algebra, and consequently does not offer any strictly new propositions. However, it does clarify certain characteristics of these two proposed measures and in so doing may assist in describing them as descriptive statistics.

Euclidean Distance

The Euclidean distance between two points in space is a well-defined concept that has been generalized to a space of any size. Providing that the space, of size k , say, has been defined by a rectangular Cartesian system of reference axes, then the square of the distance between any two points in this space is given by the sum of the squares of the differences between paired coordinates of the two points. A rectangular Cartesian system consists of k mutually perpendicular (orthogonal) axes; the pairing of coordinates is done, of course, with respect to these k reference axes. For example, suppose that four persons are located in a space of size two by the following coordinates with respect to a rectangular Cartesian system:

	Person a	Person b	Person c	Person d
Axis 1	4	0	1	3
Axis 2	9	5	8	6

The square of the distance between persons a and b , say, is given by: $(4 - 0)^2 + (9 - 5)^2 = 32$. Since squares are being summed, the result is obviously the same if we compute, instead, $(0 - 4)^2 + (5 - 9)^2 = 32$.

Designate this matrix as X . A method of securing these Euclidean distances is to operate on the matrix $X'X$, where X' is the conventional transpose of X . For these data, $X'X$ is

	a	b	c	d
a	97	45	76	66
b	45	25	40	30
c	76	40	65	51
d	66	30	51	45

The diagonal elements of $X'X$ are simply the sums of the squares of the coordinates for a given person. The off-diagonal elements are the sums of the paired products of the coordinates for the two persons designated by the row and column headings. Thus, for person a the diagonal element is $(4)^2 + (9)^2 = 97$. The element 45, occurring in row b and column a , and in row a and column b as well, is given by $(4)(0) + (9)(5) = 45$. The square of the distance between persons a and b is then given by $97 + 25 - 2(45) = 32$, as before. This in effect merely uses the principle of rewriting a square of a difference between two terms as the sum of the squares of the terms minus twice their cross-product.

The diagonal elements of the matrix $X'X$ give the squares of the lengths of each person vector in the k -space, and the off-diagonal elements give the scalar products of each pair of person vectors in this k -space. A measure of Euclidean distance between persons is thus given by the indicated operation on the matrix $X'X$, when the matrix X describes the several persons with

respect to a rectangular Cartesian system. This operation may be formulated in matrix terms. For any pair of persons, i and j , this operation consists of pre-multiplication by a row vector, E , of this form:

Persons

$$\begin{array}{cccccccccccc} a & b & \cdots & (i-1) & i & (i+1) & \cdots & (j-1) & j & (j+1) & \cdots & N \\ [0 & 0 & \cdots & 0 & +1 & 0 & \cdots & 0 & -1 & 0 & \cdots & 0] \end{array}$$

followed by post-multiplication by the transpose of this vector. For example, the square of the distance between persons a and b is given by

$$[+1 \quad -1 \quad 0 \quad 0] \cdot \begin{bmatrix} 97 & 45 & 76 & 66 \\ 45 & 25 & 40 & 30 \\ 76 & 40 & 65 & 51 \\ 66 & 30 & 51 & 45 \end{bmatrix} \cdot \begin{bmatrix} +1 \\ -1 \\ 0 \\ 0 \end{bmatrix},$$

which is equal to $(97 - 45) - (45 - 25) = 32$, as before. Thus, the square of the distance between any pair of persons is given by a product of matrices that may be written

$$EX'XE' = D^2.$$

This D^2 may be interpreted in more than one way, depending upon how one defines uncorrelated variables. This problem must now be considered.

Uncorrelated Variables

In order to show the nature of this problem, let us define, loosely, an uncorrelated form as

$$TZZ'T' = \text{a diagonal matrix,}$$

where Z is a given matrix of data and T is a transformation. To avoid discussing at this point certain problems of the form of the data, let us specify that Z consists of deviation scores that have been systematically reduced so that the variance of each row of Z equals unity. In other words, the data are taken in a form such that ZZ' is the conventional correlation matrix with units in the diagonals. Later, questions concerning the form of the data will be raised and Z will be shown to be a product of matrices, one of which is the matrix of raw scores. Now the definition of an uncorrelated form given above does not specify the non-zero elements in the diagonal matrix; in other words, it does not specify the weighting to be given each of the uncorrelated variables. Two systems of weighting appear to have special merit; one is given by

$$TZZ'T' = I$$

and is loosely related to the Mahalanobis distance measure for groups. The other is given by

$$TZZ'T' = D_{\lambda}^2,$$

where D_{λ}^2 designates the matrix of non-zero latent (or characteristic) roots of the matrix ZZ' ; this definition is associated with the Cronbach-Gleser D^2 and, as they point out, with the Pearson CRL. These two weighting systems give different results; one weights the uncorrelated variables, i.e., the factor scores, equally; the other weights the factor scores in proportion to the size of the square roots of the latent roots.

The Inverse Transformation

First consider the transformation that yields equally weighted uncorrelated variables. It always is possible to resolve Z into a product of principal-axis factors and factor scores; thus

$$Z = GD_{\lambda}P',$$

where G is a set of orthogonal columns constituting the characteristic vectors (in standard form) corresponding to the non-zero latent roots of ZZ' , D_{λ} is the matrix of positive square roots of the non-zero latent roots of ZZ' , and P' is the set of factor scores with unit variance. There now are available these generalities: $G'G = I_r$, where r is the rank of Z , and $P'P = I_r$. GG' is a pre-multiplication unit for Z and consequently a right and left unit for ZZ' ; this is true regardless of the rank of Z . Similarly, PP' is a right and left unit for $Z'Z$, and in fact the columns of P are the characteristic vectors of $Z'Z$ corresponding to the non-zero roots of $Z'Z$, which necessarily are the same as the non-zero roots of ZZ' . For a summary, see Harris (4).

These properties give a solution for T . Set

$$X = D_{\lambda}^{-1}G'Z = P'.$$

Then $XX' = I_r$, and the transformation is $T = D_{\lambda}^{-1}G'$. This principle of transformation gives as the uncorrelated data, X , the principal-axis factor scores of the persons with unit variance. With Z as defined above, these factor scores have means of zero. For the purpose of determining distances between pairs of persons this transformation leads to

$$EX'XE' = EPP'E',$$

that is, the form $X'X$ is simply the form PP' . The effect, then, of this transformation is to give distance measures that are functions of the equally-weighted principal-axis factor scores.

It is conventional to ask concerning the solution of any problem in what sense, if any, the solution is unique. Consider PP' . If $Z'Z$ is non-singular, as it might be, for example, if N , the number of persons, is less than k , the

number of variables, then PP' necessarily is simply the identity matrix, I . This means then that, using this transformation principle, studying relatively few persons with respect to relatively more linearly independent variables always yields the same numerical value for the distance between every pair of persons, regardless of what set of variables was used. If N is greater than k , then $Z'Z$ necessarily is singular and the matrix PP' is a singular idempotent matrix that is a multiplication unit for a group (in the algebraic sense) of singular matrices. This also means that there are many sets of data that will yield the same matrix, PP' , when the inverse transformation is made and the resulting $X'X$ calculated. In other words, under these conditions the distances between pairs of persons are not unique to a given set of data. For example, if we pre-multiply any given set of data, Z , by a non-singular matrix we leave invariant the matrix PP' , but not, of course, P itself. Since distance measures computed from data that have been transformed by this inverse transformation are functions of PP' , this lack of uniqueness to the given data should be recognized. It also should be recognized that these comments assume that the inverse transformation is developed from the data in hand rather than from data for a different group, such as a normative group. If the latter is done, these statements do not hold.

A direct, but quite arduous, calculation procedure would be to factor either ZZ' or $Z'Z$ in order to determine P , the matrix of factor scores. Another calculation method results from the identity

$$PP' = Z'(ZZ')^{-1}Z,$$

provided, of course, that ZZ' is non-singular. Still another calculation procedure is to utilize the principle of Rao's transformation (6) to develop a triangular matrix, C , such that

$$CZ = X.$$

Then

$$X'X = Z'C'CZ,$$

where $C'C$ is the inverse of ZZ' , provided it exists. It is interesting to observe that this latter method works even though ZZ' is singular. Adopting a new notation,

$$C'C = (ZZ')^{-1} = GD_{\lambda}^{-2}G',$$

and,

$$Z'(ZZ')^{-1}Z = PD_{\lambda}G'GD_{\lambda}^{-2}GD_{\lambda}P' = PP',$$

as before. This analysis uses the principle that if ZZ' is singular, then there exists a matrix $(ZZ')^{-1}$, which also is singular, such that

$$ZZ'(ZZ')^{-1} = (ZZ')^{-1}ZZ' = GG',$$

where GG' is the symmetric idempotent matrix that is a unit for multiplication within the group. The factored form of $(ZZ')^{-1}$ is then seen to be $GD_{\lambda}^{-2}G'$.

Principal-Axis Transformation

The inverse transformation discussed above gives as the uncorrelated form of the variables a diagonal matrix whose non-zero entries each equal unity. In other words, the inverse transformation gives uncorrelated variables of equal (unit) variance. As noted above, a different transformation may be defined by requiring that the transformation matrix, T , be such that

$$TZZ'T' = D_{\lambda}^2,$$

where, as before, D_{λ}^2 is the matrix of non-zero latent roots of ZZ' . This is the familiar canonical form of a symmetric matrix; as such, it is a well-known definition of an uncorrelated form. It differs from the inverse transformation in that the transformed variables are now weighted unequally, rather than equally, these unequal weights being given by the square roots of the roots of the characteristic equation of the symmetric matrix ZZ' . If this is chosen as the uncorrelated form, then the transformation is accomplished by setting

$$X = G'Z = D_{\lambda}P'.$$

It then follows that $XX' = D_{\lambda}^2$, as required. In order to determine distances between pairs of persons, calculate

$$EX'XE' = EPD_{\lambda}^2P'E' = EZ'ZE',$$

since GG' is a unit for multiplication, as described above, regardless of the rank of Z . In other words, choosing the canonical form of ZZ' as the uncorrelated form of the variables gives distances between pairs of persons as a function of the entries in $Z'Z$. The calculation procedure obviously requires no comment. For distance measures this solution is unique to the given set of data; this is related to the fact that the canonical form of a symmetric matrix is, under certain rather general conditions, itself unique.

It is apparent that many different diagonal matrices might be chosen as the uncorrelated form of the variables. A choice of a transformation must specify the non-zero elements of this diagonal matrix, i.e., it must specify the weights to be assigned to the variables in uncorrelated form. Two such choices that are meaningfully related to common statistical concepts are the identity matrix, I , associated with the inverse transformation, and the diagonal D_{λ}^2 , associated with the canonical form of a symmetric matrix. For both these transformations distance measures for pairs of persons are functions of factor scores; using the inverse transformation, the factor scores are weighted equally, whereas using the principal-axis transformation they are weighted unequally.

The Form of the Data

Consider now a matrix of data, Y , that consists of the observed measures, i.e., the raw scores. In order to transform these data into the form of Z ,

first write

$$YL = [y],$$

with $[y]$ the matrix of deviation scores. For any Y , which is of order k by N , the matrix L which accomplishes the transformation of raw to deviation scores is

$$\begin{bmatrix} \frac{N-1}{N} & \frac{-1}{N} & \frac{-1}{N} & \cdots & \frac{-1}{N} \\ \frac{-1}{N} & \frac{N-1}{N} & \frac{-1}{N} & \cdots & \frac{-1}{N} \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{-1}{N} & \frac{-1}{N} & \frac{-1}{N} & \cdots & \frac{N-1}{N} \end{bmatrix}$$

The matrix L is square, symmetric, of order N , and of rank $(N - 1)$. Direct multiplication verifies that $L = L^2$, i.e., that L is idempotent. The matrix L is a quadratic form with roots of unity and is an example of the type of matrix referred to in Cochran's theorem (1). Further, L is a multiplication unit for any vector consisting of N terms that sum to zero; the row vector E employed earlier is such a vector. Note that what is being done here is to take an operation that is ordinarily considered to be an additive one and to write it as a multiplicative operation; this becomes a useful tool in the analysis of certain relationships among matrices. It is now possible to write

$$SYL = Z,$$

that is, to pre-multiply the deviation scores by the appropriate diagonal matrix to secure the form Z . The diagonal matrix is, of course, one in which each of the non-zero elements is given by the reciprocal of the product of the square root of N and the standard deviation of the variables.

Using this definition of Z and noting that S is a non-singular matrix, that L is idempotent, and that L is a multiplication unit for E , the distance between any pair of persons under the inverse transformation becomes

$$EZ'(ZZ')^{-1}ZE' = EY'(YLY')^{-1}YE'.$$

(A method of calculation when ZZ' is singular was suggested above.) Reduced to these terms, then, this distance measure is a function of the raw scores and of the inverse of the matrix of variances and covariances. An analogous reduction of the Cronbach-Gleser measure gives

$$EZ'ZE' = EY'S^2YE',$$

showing that it is a function of the raw scores and the reciprocals of the variances of the variables, since the scalar $1/N$ affects all pairs in the same way. Clearly, the two measures are identical if, and only if, YLY' is a diagonal matrix of variances. It also is evident that

$$EY'S^2YE' \neq EY'YE',$$

that is, that any change in scale for one or more variables affects the Cronbach-Gleser measure. This is not true for the measure derived from the inverse transformation.

Other modifications of the form of the data might be explored using these techniques. One such modification that is of interest is secured by centering Y by columns rather than by rows; we may write MY to designate this, where M is of the form of L with k substituted for N . The matrix MY is such that each column sums to zero, i.e., the "elevation" has been equalized for all persons. Then by choosing the appropriate diagonal matrix, S_N , the product $S_N Y' M Y S_N$ yields the intercorrelations of the persons. If the principal-axis factors of this matrix are taken as the descriptions of persons in terms of uncorrelated variables, then the distance between any pair of persons is simply $E S_N Y' M Y S_N E'$. Obviously,

$$E S_N Y' M Y S_N E' \neq E Y' M Y E',$$

which merely states that any change in scale for the columns affects this type of measure.

Finally, MYL is a double-centered matrix, i.e., it sums to zero both by rows and by columns. Since matrix algebra is a linear associative algebra, it makes no difference whether one first forms MY and then centers by rows, or first forms YL and then centers by columns; the resulting MYL is the same in either case. Now, since L is a multiplication unit for any E , it can be seen that

$$E Y' M Y E' = E L Y' M Y L E'.$$

This identity emphasizes the point that if the data are centered by columns then double-centering does not alter this particular distance measure.

Table 1 gives the algebra of distance measures developed on the basis of the two principles discussed above for various forms of the data. It is evident that not all the distance measures listed there would be judged to be meaningful ones; it also is evident that some of them are duplicates. The table is probably of primary value in demonstrating that this system of analysis makes explicit a range of choices of distance measures and provides a method of specifying the choice that a worker may make. Table 2 gives illustrative scores on three tests for five students. These scores have been used to compute the illustrative distance measures between pairs of students that are given in Table 3.

TABLE 1

Distance Measures for Various Forms of the Data

Form of Data	Principal-axis Transformation	Inverse Transformation
Y	$EY'YE'$	$EY'(YY')^{-1}YE'$
YL	$EY'YE'$	$EY'(YLY')^{-1}YE'$
SYL	$EY'S^2YE'$	$EY'(YLY')^{-1}YE'$
MY	$EY'MYE'$	$EY'M(MYY'M)^{-1}MYE'$
MYS_N	$ES_NY'MYS_N'E'$	$ES_NY'M(MYS_N'Y'M)^{-1}MYS_N'E'$
MYL	$EY'MYE'$	$EY'M(MYLY'M)^{-1}MYE'$

TABLE 2

Illustrative Distance Measures between Pairs of Students

TABLE 2

Illustrative Scores on Three Tests for Five Students

	Student				
	a	b	c	d	e
1. Spelling	18	21	11	16	15
2. Usage	44	43	39	51	46
3. Vocabulary	18	17	14	14	20

$EY'YE'$					$EY'(YY')^{-1}YE'$				
	a	b	c	d		a	b	c	d
b	11				b	.32			
c	90	125			c	.52	1.46		
d	69	98	169		d	1.22	1.54	.73	
e	17	54	101	62	e	.47	1.56	.32	1.88

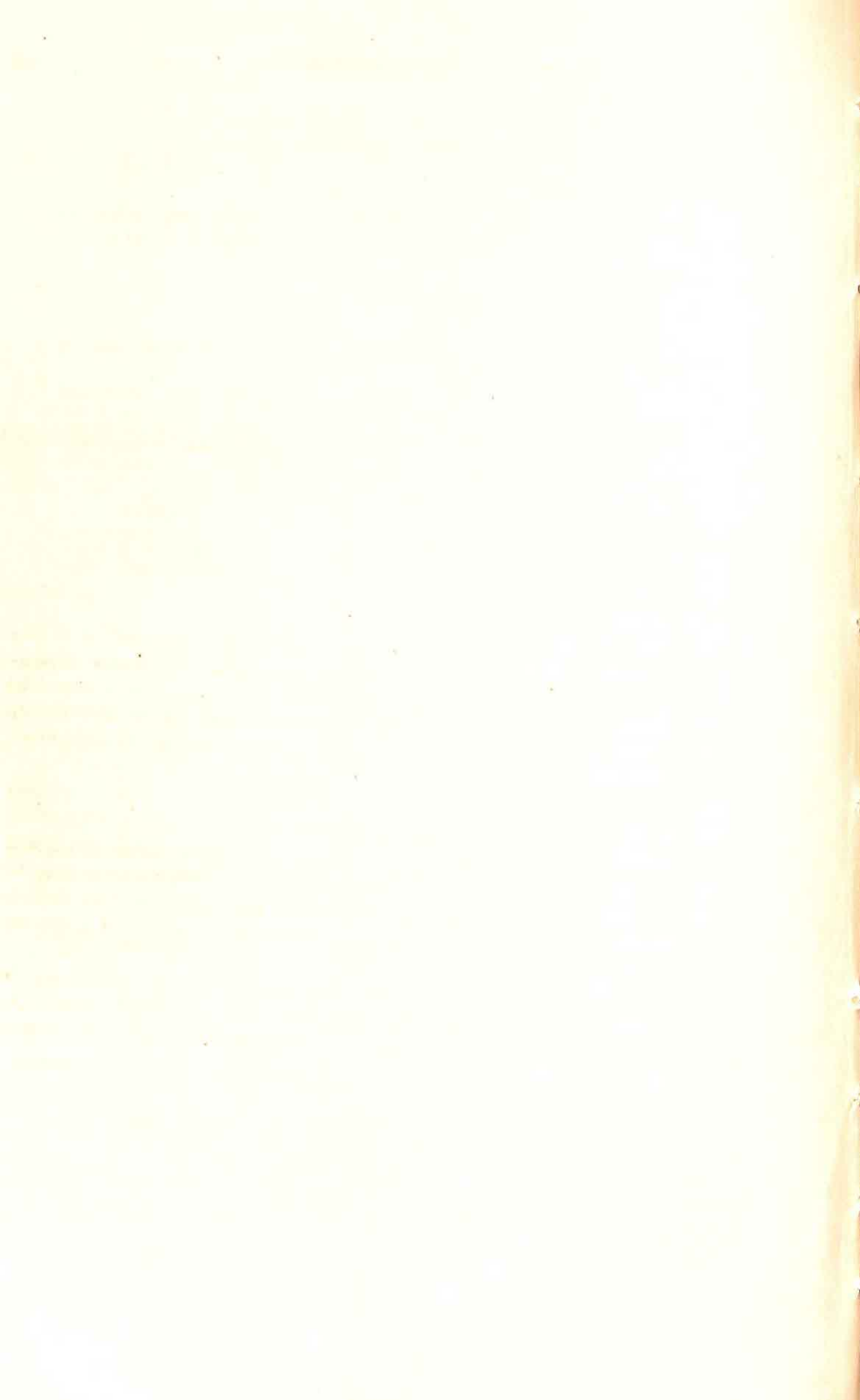
$EY'S^2YE'$					$EY'(YLY')^{-1}YE'$				
	a	b	c	d		a	b	c	d
b	1.07				b	.37			
c	9.03	11.82			c	1.23	1.84		
d	6.48	8.08	11.61		d	1.30	1.80	2.00	
e	1.82	5.52	11.25	8.33	e	.65	1.99	1.92	1.89

REFERENCES

1. Cochran, W. G. The distribution of quadratic forms in a normal system with applications to the analysis of variance. *Proc. Cambridge Phil. Soc.*, 1934, 30, 178-91.
2. Cronbach, L. J., and Gleser, G. C. Assessing similarity between profiles. *Psychol. Bull.*, 1953, 50, 456-73.
3. Gaier, E. L., and Lee, M. C. Pattern analysis: the configural approach to predictive measurement. *Psychol. Bull.*, 1953, 50, 140-48.
4. Harris, Chester W. Relations among factors of raw, deviation, and double-centered score matrices. *J. exp. Educ.*, 1953, 22, 53-58.
5. Osgood, C. E., and Suci, G. J. A measure of relation determined by both mean difference and profile information. *Psychol. Bull.*, 1952, 49, 251-62.
6. Rao, C. R. Advanced statistical methods in biometric research. New York: Wiley, 1952.
7. Thorndike, R. L. Who belongs in the family? *Psychometrika*, 1953, 18, 267-76.
8. Webster, H. A note on profile similarity. *Psychol. Bull.*, 1952, 49, 538-39.

Manuscript received 4/19/54

Revised manuscript received 11/8/54



THE ESTIMATION OF THE DISCRIMINAL DISPERSION IN THE METHOD OF SUCCESSIVE INTERVALS*

RAYMOND H. BURROS

INSTITUTE FOR MOTIVATIONAL RESEARCH, INC., CROTON-ON-HUDSON, N. Y.

A new algebraic formula is derived for estimation of the discriminial dispersion in the method of successive intervals. The legitimate use of the formula requires that as many normal deviates as possible be present in the matrix. For this reason, it is recommended that deviates corresponding to the interval (0.01, 0.99) of the cumulative proportions be used, instead of those corresponding to (0.05, 0.95), the interval used by Edwards and Thurstone. Computations on data published by Edwards and Thurstone showed that when adjustment was made for variability in dispersions calculated by the formula of this paper, a reduction of fifty per cent in mean absolute discrepancy was produced. Since the formula is easy to use and avoids the disadvantages of its predecessors, it should have fairly wide applicability in psychological research.

The method of successive intervals is perhaps the most practical way of obtaining *rational* scale values of stimuli along a unidimensional psychological continuum not simply correlated with any physical variable. The data may be provided by any procedure in which judges classify stimuli into a finite number of mutually exclusive and exhaustive classes which are ordered along some dimension.

When the number of stimuli is small, they may be ranked without ties, so that the number of classes equals the number of stimuli. When the number of stimuli is large, they may be either sorted into piles or rated on a rating scale. With either of these procedures, the number of classes may be considerably less than the number of stimuli. For adequate reliability, a large sample of judges is needed when any of these techniques of gathering data is used.

Although successive intervals was developed by L. L. Thurstone, its first published account was given in a paper by Saffir (8) in 1937. Recently, papers by Edwards (4) and Edwards and Thurstone (5) have presented a

*This research was supported in part by the United States Air Force under Contract No. AF 33(038)-25726 monitored by Air Force Personnel and Training Research Center. Permission is granted for reproduction, translation, publication, use and disposal in whole and in part by or for the United States Government. The writer is grateful to Dr. A. L. Edwards for a critical reading of an earlier version of this paper, and to Dr. L. H. Lanier and Dr. L. M. Stolurow for editorial advice on the present version, which was written at the University of Illinois. The editors of *Psychometrika* have informed the writer that H. J. A. Rimoldi and M. Hormaeche (7) have independently derived the same formula for the discriminial dispersion from a different set of postulates using the law of comparative judgment.

check on internal consistency which indirectly tests the applicability of the postulates to any particular set of data. This check now makes successive intervals a serious rival to the method of paired comparisons. The advantage of successive intervals over paired comparisons lies in its greater speed in collecting data. Empirical studies (4, 5) have shown that there is a linear relation between scale values obtained by these two methods.

In any stimulus scaling method developed in the Thurstone manner, there are at least two important kinds of parameters, represented respectively by S_j , the scale value of the j th stimulus, and σ_j , the corresponding discriminial dispersion. Although adequate computational techniques for estimating each S_j by the method of successive intervals have been published (4, 5), those available for estimating σ_j are subject to improvement.

The first technique, developed by Thurstone and presented by Saffir (8), does not base the computation of each σ_j on all of the data. Also, it does not use a simple algebraic formula in the manner originated by Thurstone (9, 10) and further applied by Burros (2) and Burros and Gibson (3) for estimation of σ_j in the method of paired comparisons. It is interesting to note, therefore, that in a recent paper on successive intervals, Edwards and Thurstone (5) did not use the technique presented by Saffir for estimating the dispersions. Instead these writers used one published by Attneave (1). A critical examination of Attneave's technique will be made later on in this paper. In the writer's opinion, it does not have a rigorous basis.

Perhaps the most rigorous approach to the problem is a least squares solution recently published by Gulliksen (6). Unfortunately, it is possible (although admittedly improbable) that negative estimates of the dispersions may be calculated by this technique. This sort of result could happen if the dispersions are exceedingly variable. A small positive dispersion could then be estimated as negative when his least squares solution is applied to the data. A related discussion of this problem of absurd results in paired comparisons is presented by Burros and Gibson (3, pp. 63-64).

Since the techniques published by Saffir (8) and Attneave (1) are questionable, and the one by Gulliksen (6) conceivably may give absurd results, a new formula may be of interest. This paper, therefore, presents the derivation of a simple formula for the estimation of σ_j in the method of successive intervals, which is similar to those previously derived for paired comparisons (10, 2, 3). The use of the formula will then be illustrated by means of further analysis of data presented by Edwards and Thurstone (5).

Definition of Symbols

R = postulated unidimensional psychological continuum with finite range arbitrarily divided into N class intervals corresponding to the steps on an N -point rating scale;

- S \equiv postulated unidimensional psychological continuum with unrestricted range corresponding to R ;
 R'_k \equiv upper true limit of k th class interval of R ;
 S'_k \equiv corresponding upper true limit of k th class interval of S ;
 \dot{R}_j \equiv momentary estimate by a judge of the scale position of the j th object on R ;
 \dot{S}_j \equiv corresponding momentary estimate of the scale position of the j th object on S ;
 S_j \equiv scale position (mean and median of \dot{S}_j) of j th object on S ;
 σ_j \equiv disciminal dispersion or standard deviation of distribution of \dot{S}_j ;
 X_{jk} \equiv $(S'_k - S_j)/\sigma_j$;
 \dot{z}_j \equiv $(\dot{S}_j - S_j)/\sigma_j$;
 P_{jk} \equiv probability that $\dot{R}_j \leq R'_k$.

Postulates

1. There exists a unidimensional psychological continuum (R) with a finite range, arbitrary units, and an arbitrary origin.
2. There exists a corresponding unidimensional psychological continuum (S) with an unrestricted range, equal units, and an arbitrary origin.
3. $S = f(R)$, where the function is monotonic, increasing, and generally nonlinear.
4. For object j , and corresponding to each observed momentary estimate (\dot{R}_j) by a judge on R , there exists a theoretical momentary estimate (\dot{S}_j) on S . The distribution of \dot{S}_j is normal with mean (and thus median) of S_j and standard deviation σ_j .

Basic Theorem

Since $f(R)$ is monotonic increasing,

$$P_{jk} \equiv P(\dot{R}_j \leq R'_k) = P(\dot{S}_j \leq S'_k).$$

But if $\dot{S}_j \leq S'_k$, then

$$\dot{S}_j - S_j \leq S'_k - S_j \quad \text{and} \quad (\dot{S}_j - S_j)/\sigma_j \leq (S'_k - S_j)/\sigma_j,$$

so that

$$\dot{z}_j \leq X_{jk}$$

by definition of these quantities. Therefore,

$$P_{jk} = P(\dot{z}_j \leq X_{jk}) = G(X_{jk}),$$

where G is the normal probability integral. Given each estimated value of P_{jk} determined by the empirical frequency distribution of \dot{R}_j on R , therefore, the corresponding estimated value of X_{jk} can be found from a table of the normal integral. These may be arranged in a matrix X .

Whenever P_{jk} equals 0 or 1, the value of X_{jk} is indeterminate. If any proportion is too near to either 0 or 1 (say, less than 0.01, or greater than 0.99) the values of X_{jk} are too unreliable to be recorded. Whenever a value of X_{jk} is indeterminate or unreliable, it is omitted from the X matrix.

Derivation of Formula for σ_i

From the definition of X_{jk} , it follows that

$$S'_k = S_j + \sigma_j X_{jk}. \quad (1)$$

Similarly for X_{ik} ,

$$S'_k = S_i + \sigma_i X_{ik}. \quad (2)$$

Therefore,

$$S_j + \sigma_j X_{jk} = S_i + \sigma_i X_{ik} \quad (3)$$

and

$$X_{jk} = (S_j - S_i)/\sigma_j + (\sigma_i/\sigma_j)X_{ik}. \quad (4)$$

Equation (4) says that the j th row in the X matrix is theoretically a linear function of the i th row with slope of

$$m = \sigma_i/\sigma_j. \quad (5)$$

Theoretically these two rows are perfectly correlated.

Let V_i and V_j be the respective measures of variability, e.g., standard deviations or ranges, of the i th and j th rows of X . Assuming perfect correlation, therefore, the slope of (4) is also equal to

$$m = V_j/V_i. \quad (6)$$

Therefore,

$$\sigma_i/\sigma_j = V_j/V_i \quad (7)$$

and

$$\sigma_i V_i = \sigma_j V_j. \quad (8)$$

Thus, for any two stimulus objects i and j either side of (8) theoretically equals a constant, defined as

$$\alpha = \sigma_i V_i. \quad (9)$$

Therefore,

$$\sigma_i = \alpha/V_i. \quad (10)$$

In order to estimate α , a unit of σ_i must be chosen. This is an arbitrary matter. The simplest definition is that the unit is the mean of the sigmas of the n stimuli, i.e.,

$$(\sum_i \sigma_i)/n \equiv 1 \quad (11)$$

and

$$\sum_i \sigma_i = n. \quad (12)$$

Summing (10), and using (12),

$$n = \sum_i \sigma_i = \alpha \sum_i (1/V_i). \quad (13)$$

Therefore,

$$\alpha = n / \sum_i (1/V_i). \quad (14)$$

Thus, α is the harmonic mean of the values of V_i , which are obtained empirically from the rows of the X matrix. After α is estimated from (14), each estimated value of σ_i is given by (10).

Now that the new formula has been derived, it is possible to criticize Attneave's technique. According to Attneave, this "assumes that the mean dispersion of stimuli represented in one dichotomy is equal to the mean dispersion of those represented in another; this assumption may be only approximately correct" (1, p. 340). A sufficient condition for this assumption is that all entries in the X matrix are present. When this is so, it follows from Attneave's directions that the discriminial dispersion of any stimulus equals the ratio of the *arithmetic* mean of the ranges of stimulus X values to the range of the given stimulus. Equations (10) and (14) show, however, that the proper average for the numerator of the ratio is the *harmonic* mean, *not* the *arithmetic* mean. Even when Attneave's assumption is known to be true, therefore, his technique is not strictly correct. It may sometimes give adequate results, however, if the arithmetic mean range and the corresponding harmonic mean are approximately equal.

Application of Formula

To save space, tables presented by Edwards and Thurstone (5) will not be reproduced. They provide the following relevant data: (a) their Table 1 (5, p. 172) gives the cumulative proportions P_{ik} for ten stimuli rated on a nine-point scale; (b) their Table 2 (5, p. 173) gives the normal deviates X_{ik} corresponding to the proportions in the closed interval (0.05, 0.95). In order to reduce the number of empty cells, the writer entered into a copy of this table those additional deviates required to encompass the interval (0.01, 0.99) of the proportions.

Table 1 of this paper shows the results of the computations. V_j is the standard deviation of the reliable normal deviates for stimulus j based upon N_j values. The parameter α is then computed by (14) to be

$$\alpha = n / \sum_j (1/V_j) = 10/8.332 = 1.20.$$

Then each value of σ_j is computed in Table 1 as

$$\sigma_j = \alpha(1/V_j).$$

TABLE 1
Calculation of the Discriminal Dispersion (σ_j) from Data
Presented by Edwards and Thurstone (5)

j^*	N_j^{**}	V_j	$1/V_j$	σ_j
1	6	1.08	0.926	1.11
2	7	1.26	0.794	0.95
3	6	1.16	0.862	1.03
4	8	1.25	0.800	0.96
5	8	1.24	0.806	0.97
6	6	1.52	0.658	0.79
7	8	1.34	0.746	0.90
8	7	1.48	0.676	0.81
9	8	0.998	1.012	1.21
10	8	0.950	1.052	1.26
Sum			8.332	9.99***

*As rank order (j) of the stimulus increases, the scale value tends to decrease.

** N_j is the number of normal deviates corresponding to proportions in the interval (0.01, 0.99).

***Presumably errors from rounding off decimals account for the departure of this sum from the theoretical value (10.00).

Discussion

The estimates of the dispersions calculated for successive intervals by the formula of this paper correspond roughly to the estimates of the same parameters of the same stimuli computed by Edwards and Thurstone (5, p. 177) by means of the method of paired comparisons. They reported that the latter dispersions "showed considerable variation, ranging from a low of .52 for stimulus 6 to a high of 1.32 for stimulus 10." (5, p. 177). The corresponding successive intervals dispersions in Table 1 of this paper are 0.79 and 1.26, respectively.

Although Edwards and Thurstone (5, p. 177) reported comparable variation in σ_j computed from their successive intervals data by Attneave's technique (1), they noted the surprising fact that adjustment for variability of dispersions did not improve the goodness of fit measured by the mean absolute discrepancy of the proportions. Since this does not conform to usual

experience with paired comparisons, some additional computations made by the writer may be of interest.

First, using the normal deviates corresponding to proportions in the interval (0.01, 0.99) and the dispersions previously computed by the new formula, the successive intervals scale values of these stimuli were computed by an algebraic technique. Since there was close correspondence with the scale values reported by Edwards and Thurstone, who used the interval (0.05, 0.95), no details about these computations need be given here.

Then the mean absolute discrepancy was computed to be 0.0135. This is half of the value, 0.027, reported by Edwards and Thurstone when adjustment was made in successive intervals for variability in dispersions calculated by Attneave's technique. It is concluded that the mean absolute discrepancy may be considerably less than that reported by Edwards and Thurstone when correction is made for variability in dispersions calculated by the formula of this paper.

In order to fulfill the requirements of the formula, however, the number of empty positions in the X matrix should be reduced. The use of a wider interval of acceptably reliable proportions, i.e., (0.01, 0.99) instead of (0.05, 0.95) will produce this desired result. The use of this wider interval is, therefore, recommended.

Since the formula presented here avoids the disadvantages of its predecessors but is easy to use, it should have fairly wide applicability in psychological research.

REFERENCES

1. Attneave, F. A method of graded dichotomies for the scaling of judgments. *Psychol. Rev.*, 1949, **56**, 334-340.
2. Burros, R. H. The application of the method of paired comparisons to the study of reaction potential. *Psychol. Rev.*, 1951, **58**, 60-66.
3. Burros, R. H., and Gibson, W. A. A solution for Case III of the law of comparative judgment. *Psychometrika*, 1954, **19**, 57-64.
4. Edwards, A. L. The scaling of stimuli by the method of successive intervals. *J. appl. Psychol.*, 1952, **36**, 118-122.
5. Edwards, A. L., and Thurstone, L. L. An internal consistency check for scale values determined by the method of successive intervals. *Psychometrika*, 1952, **17**, 169-180.
6. Gulliksen, H. A least squares solution for successive intervals. *American Psychologist*, 1952, **7**, 408 (abstract).
7. Rimoldi, H. J. A., and Hormaeche, M. The law of comparative judgment in the successive intervals and graphic rating scale methods. *Psychometrika*, 1955, **20**, 307-318.
8. Saffir, M. A. A comparative study of scales constructed by three psychophysical methods. *Psychometrika*, 1937, **2**, 179-198.
9. Thurstone, L. L. A law of comparative judgment. *Psychol. Rev.*, 1927, **34**, 273-286.
10. Thurstone, L. L. Stimulus dispersions in the method of constant stimuli, *J. exp. Psychol.*, 1932, **15**, 284-297.

Manuscript received 6/25/54

Revised manuscript received 12/2/54

THE LAW OF COMPARATIVE JUDGMENT IN THE SUCCESSIVE INTERVALS AND GRAPHIC RATING SCALE METHODS*

H. J. A. RIMOLDI†

UNIVERSITY OF CHICAGO

AND

M. HORMAECHE

UNIVERSITY OF VIRGINIA

The law of comparative judgment is applied to the successive intervals and graphic rating scale methods. A procedure for estimating the modal discriminial process and discriminial dispersion of the stimuli, as well as the value of the boundaries of the intervals on the continuum, is given. From the estimated values it is possible to determine the theoretical proportions and to compare them with the actual experimental proportions. The agreement between these values is an indication of the adequacy of the assumptions made.

The rationale and the system of computations described in the present paper developed from a suggestion offered by L. L. Thurstone in one of his courses at the University of Chicago. He suggested an interpretation of the method of successive intervals based on the assumption that, in the process of indicating preferences, a subject will compare the affective value of each stimulus with the affective value represented by the interval limits on the psychological continuum.

In the present study stimuli were presented using three different procedures:

1. *Method of successive intervals.* The subject was presented with equally spaced intervals having reference to degree of interest in an indicated stimulus. His placement of a check mark in any interval was interpreted as indicating that his interest in that stimulus was greater than that represented by the lower limit of the interval and smaller than the interest represented by its upper limit. A pre-test on approximately 30 subjects demonstrated experimentally that the continuum could be defined unambiguously.

*This article is the first part of a larger study conducted at the *Laboratorio de Psicología, Facultad de Humanidades y Ciencias*, Montevideo, Uruguay, during the years 1951 and 1952. The authors want to thank Dr. L. V. Jones for his critical comments on the manuscript. The authors have been informed by the editors of *Psychometrika* that R. H. Burros (2) has independently reached the same analytic solution for the computation of stimulus dispersions. Dr. Burros has used a set of assumptions different from the ones stated in the present paper.

†Now at Loyola University, Chicago, Ill.

2. *Multiple category method.* This is a variation of the previous procedure. Subjects were instructed to encircle the word or sign (*Yes, yes, ?, no, No*) that best represented their interest in the stimulus.

3. *Graphic rating scale.* Here the subject was asked to state his interest in each stimulus by placing a check mark on a straight line without intervals. The location of the check mark on the continuum was interpreted as indicating that the interest of the subject in the stimulus was greater than that represented by the points on the continuum located to the left of the check mark, and smaller than that represented by the points on the continuum located to the right of the check mark.

In all the presentations it was assumed that the subject compared the value of the stimulus with the value represented by the different points on the continuum.

Determination of L_i

Let S_j ($j = 1, 2, \dots, j, \dots, n$) represent the modal discriminial process for the j th stimulus, and σ_j ($j = 1, 2, \dots, j, \dots, n$) the discriminial dispersion for the j th stimulus. L_i ($i = 1, 2, \dots, i, \dots, m - 1$) is the modal discriminial process of the boundary between intervals i and $i + 1$, where there are m successive intervals, and d_i ($i = 1, 2, \dots, i, \dots, m - 1$) represents the discriminial dispersion of the i th boundary. d_i is generated in a manner similar to that described by Thurstone (7) for the discriminial dispersion of the stimuli.

It will be assumed that the stimuli are normally distributed and that, together with the interval limits, they can be located on the same psychological continuum. Throughout the study it will be assumed that we are dealing with Thurstone's case II (8), where several individuals made one judgment each.

The origin of the scale will be defined as

$$\sum_i S_i = 0, \quad (1)$$

and the unit of measurement will be

$$(\sum_i \sigma_i)/n = 1. \quad (2)$$

According to the law of comparative judgment (8) and to the previous assumptions it is possible to write

$$L_i - S_j = X_{ij} \sqrt{d_i^2 + \sigma_j^2 - 2r_{ij}\sigma_j d_i}, \quad (3)$$

where X_{ij} is the normal deviate corresponding to the proportion of times that stimulus j has been placed in a position less preferred than the point L_i . r_{ij} is the correlation between the subject's judgment of stimulus j and the interval boundary i .

It seems defensible to assume that d_i will be very small when compared

with σ_i and that the more precise the definition of the continuum the smaller the value of d_i . This assumption seems to be corroborated by an unpublished investigation by L. V. Jones at the University of Chicago. The empirical evaluation of d_i , in terms of the unit of measurement of (2), demonstrates the magnitude to be no greater than .05 and generally much smaller. Ignoring the value of d_i , (3) becomes

$$L_i - S_i = X_{ii}\sigma_i. \quad (4)$$

Adding and averaging (4) for all stimuli and keeping L_i constant we have, using (1),

$$(\sum_i X_{ii}\sigma_i)/n = L_i. \quad (5)$$

Determination of the Modal Discriminal Process for Each Stimulus

According to (4) it is possible to have as many S_i values as there are L_i points on the continuum. Keeping S_i constant and adding and averaging for all the L_i values, we obtain

$$(\sum_i L_i - \sigma_i \sum_i X_{ii})/(m-1) = S_i. \quad (6)$$

Determination of the Discriminal Dispersions

Subtracting (4) for stimuli 1 and 2, we have

$$\begin{aligned} L_i - S_1 &= X_{i1}\sigma_1 \\ L_i - S_2 &= X_{i2}\sigma_2 \\ \hline S_2 - S_1 &= X_{i1}\sigma_1 - X_{i2}\sigma_2. \end{aligned} \quad (7)$$

Equation (7) is similar to the basic equation used in the scaling of mental tests (6) and may be written

$X_{i1} = X_{i2}(\sigma_2/\sigma_1) + (S_2 - S_1)/\sigma_1$, where $(S_2 - S_1)/\sigma_1 = K = \text{constant}$.

Thus,

$$X_{i1} = X_{i2}(\sigma_2/\sigma_1) + K. \quad (8)$$

There are as many X_{i1} , X_{i2} , \dots , X_{ii} , \dots , X_{in} values as there are i points on the continuum. From these values n standard deviations, V_i , may be computed by

$$V_i = \sqrt{(m-1) \sum_i X_{ii}^2 - (\sum_i X_{ii})^2 / (m-1)}.$$

Performing the necessary operations, (8) can be written as

$$V_1 = V_2(\sigma_2/\sigma_1)$$

and consequently

$$\sigma_2 = c/V_2, \quad \text{where} \quad V_2\sigma_2 = c = \text{constant}. \quad (9)$$

Changing subscripts of σ and V from 1 to j and adding all the resulting

equations,

$$n = c \sum_i (1/V_i) \quad \text{since} \quad \sum_i \sigma_i = n,$$

and accordingly

$$c = n / \sum_i (1/V_i). \quad (10)$$

From (9) and (10) it is readily seen that the value of the disciminal dispersion for any stimulus j is given by

$$\sigma_i = n / [V_i \sum_i (1/V_i)]. \quad (11)$$

If the values of the dispersions are equal for all the stimuli, then, plotting (8), the slope of the line will be unity. If it can be safely assumed that the disciminal dispersions for all the stimuli are equal, then according to (2) their value should be unity. Consequently (5) and (6) can be written as

$$L_i = \sum_i X_{ii}/n, \quad (12)$$

and

$$S_i = (\sum_i L_i - \sum_i X_{ii}) / (m - 1). \quad (13)$$

Reproduction of the Original Experimental Proportions

From (4)

$$X_{ii} = (L_i - S_i) / \sigma_i. \quad (14)$$

If the disciminal dispersions are assumed to be equal, then

$$X_{ii} = (L_i - S_i). \quad (15)$$

The X_{ii} values thus obtained should be transformed into proportions and these compared with the original experimental values.

Equations (1) to (15) refer to true values. For the purposes of computation a parallel set of equations can be written using sample values instead of true values.

Experimental Results

Subjects were asked to state their degree of "interest in knowing" certain people. The stimuli were names of men and women who were well known to the experimental group. (See Tables 2, 3, 4, and 5).

TABLE 1

Stimuli	Known Values of σ_j	Values of σ_j from Formula (11)
a	.39	.40
b	1.51	1.53
c	.43	.43
d	1.10	1.06
e	.64	.64
f	1.34	1.38
g	1.60	1.55

TABLE 2

Method of Successive Intervals

(Decimal points omitted.)

		1	2	3	4	5	6	7	8	9
Roosevelt	A*	000	006	000	041	065	176	206	318	183
	T*	005	007	015	027	086	162	186	294	218
Leonardo	A	006	012	024	035	076	206	159	259	224
	T	009	011	020	034	096	163	174	274	218
Hitler	A	076	047	029	059	100	141	147	247	153
	T	070	038	048	059	122	151	134	188	190
Garibaldi	A	006	053	064	171	265	276	076	065	024
	T	034	045	077	112	276	211	143	090	002
Marie Antoinette	A	018	065	100	065	194	147	141	176	094
	T	040	038	056	075	169	209	165	174	074
Mussolini	A	047	094	094	088	206	159	118	159	035
	T	070	055	078	099	190	210	137	123	038
Joan of Arc	A	012	012	012	053	088	235	182	229	176
	T	010	013	025	039	111	184	185	264	169
Shakespeare	A	006	012	018	006	053	129	212	288	276
	T	007	009	015	025	073	129	155	271	316
San Martín	A	012	018	047	065	200	253	194	188	024
	T	011	021	042	071	196	273	196	161	029
Becquer	A	012	076	094	059	159	176	147	153	124
	T	038	033	052	069	164	208	165	187	084
Isabella	A	041	065	094	082	312	188	141	041	035
	T	054	055	083	110	222	228	135	095	018
Cleopatra	A	059	035	070	065	176	188	153	159	094
	T	058	044	059	075	162	189	149	172	092
Dante	A	035	024	029	041	129	159	182	276	124
	T	028	025	056	036	129	178	166	226	156
Beethoven	A	000	018	024	024	065	141	159	300	270
	T	009	010	019	028	079	139	156	269	291
Cervantes	A	000	006	024	065	176	312	224	188	006
	T	005	013	027	054	169	272	224	197	039
Napoleon	A	006	012	012	047	076	106	153	294	294
	T	009	010	018	033	072	132	155	262	309
Chopin	A	018	018	047	029	094	182	159	229	224
	T	020	019	031	041	107	160	162	242	218
Michelangelo	A	012	012	012	041	053	165	194	288	224
	T	010	011	021	032	090	152	164	272	248
Columbus	A	018	018	047	047	188	282	188	141	070
	T	017	023	042	067	174	237	192	191	057
Gandhi	A	041	053	041	041	218	141	135	224	106
	T	048	034	052	066	152	188	155	192	113
Ruben Darío	A	024	012	047	082	170	170	188	212	094
	T	022	025	042	060	153	214	182	208	094
Madame Curie	A	006	012	006	035	100	170	188	335	147
	T	006	010	018	033	099	178	196	281	179
Pasteur	A	006	000	000	018	047	176	188	412	153
	T	002	003	008	019	070	159	203	330	206
Sarmiento	A	053	041	065	076	188	200	165	159	053
	T	051	044	064	083	183	208	155	155	057
Bolívar	A	006	012	035	106	229	224	165	170	053
	T	010	018	038	063	183	263	204	182	039

Average Discrepancy (per column) 007 009 012 014 021 029 020 029 019

Standard Deviation = .025

Average Discrepancy (total) = .017

*A = Actual proportions; T = Theoretical proportions.

TABLE 3

Graphic Rating Scale

(Decimal points omitted.)

		1	2	3	4	5	6	7	8	9
Roosevelt	A*	006	006	012	018	147	229	241	129	212
	T*	005	008	013	024	168	195	197	187	203
Leonardo	A	006	012	012	024	229	147	200	200	170
	T	008	011	017	030	189	193	192	176	184
Hitler	A	076	047	029	035	194	135	176	182	123
	T	070	036	041	056	218	154	134	120	171
Garibaldi	A	035	041	041	094	447	194	070	065	012
	T	034	040	060	093	383	219	109	048	014
Marie Antoinette	A	070	053	070	065	300	165	112	094	070
	T	078	047	059	077	279	176	125	088	071
Mussolini	A	118	088	053	112	270	129	112	076	041
	T	129	063	072	088	281	155	101	064	047
Joan of Arc	A	018	024	018	041	265	194	165	147	129
	T	018	020	028	045	237	204	177	144	127
Shakespeare	A	000	012	024	012	141	218	194	176	223
	T	002	003	007	015	147	204	224	211	187
San Martín	A	012	012	041	094	294	212	194	082	059
	T	014	019	031	057	308	245	172	104	050
Becquer	A	035	053	082	053	259	200	129	094	094
	T	052	037	049	068	274	190	143	103	084
Isabella	A	053	088	065	112	406	141	100	024	012
	T	066	059	081	113	379	179	081	033	009
Cleopatra	A	070	082	053	100	276	129	135	070	082
	T	092	050	061	075	274	167	120	086	075
Dante	A	018	018	024	029	259	188	223	153	088
	T	016	017	028	045	250	219	180	139	106
Beethoven	A	018	012	029	018	100	159	129	223	312
	T	016	012	018	026	143	141	148	166	330
Cervantes	A	006	000	018	041	206	223	259	123	123
	T	006	012	014	031	218	227	208	165	119
Napoleon	A	000	018	012	029	159	123	147	165	347
	T	010	010	014	023	135	145	155	174	334
Chopin	A	029	012	041	024	159	194	147	170	223
	T	026	020	026	039	184	161	154	151	239
Michelangelo	A	018	006	018	018	141	176	223	159	241
	T	013	011	019	029	164	165	166	172	261
Columbus	A	029	018	041	088	388	194	106	065	070
	T	028	030	046	067	317	221	151	092	048
Gandhi	A	041	065	041	076	270	165	100	123	118
	T	057	038	047	064	254	180	139	112	109
Ruben Darío	A	029	041	041	041	265	241	165	100	076
	T	032	030	040	062	284	208	154	112	078
Madame Curie	A	006	012	012	076	212	218	153	159	153
	T	010	012	021	036	219	210	194	160	138
Pasteur	A	006	000	000	047	123	270	206	153	194
	T	007	010	015	027	177	193	189	181	201
Sarmiento	A	059	059	053	070	282	188	100	106	082
	T	068	045	053	073	273	179	133	095	081
Bolívar	A	024	035	041	065	274	206	100	141	112
	T	030	027	038	054	260	201	160	124	106
Average Discrepancy (per column)		006	010	010	014	024	024	028	021	012
Standard Deviation = .022										
Average Discrepancy (total) = .016										

*A = Actual Value; T = Theoretical Value.

TABLE 5
Paired Comparisons

	Beethoven	Michelangelo	Napoleon	Roosevelt	Chopin	Joan of Arc	Cervantes	Hitler	Bolívar	Becquer	San Martín	Columbus	Sarmiento	Garibaldi	Mussolini
Beethoven															
Michelangelo	615*														
Napoleon	579**	574	562												
Roosevelt		615	704	527											
Chopin		674	618	599											
Joan of Arc		645	527	485	509										
Cervantes		614	552	508	464										
Hitler		728	645	527	320	527									
Bolívar		663	610	575	516	540									
Becquer		769	734	598	444	609	562								
San Martín		709	670	666	606	591	564								
Columbus		675	657	728	698	580	580								
Sarmiento		722	691	698	640	610	587	528							
Garibaldi		716	669	686	769	604	586	538	532						
Mussolini		763	710	692	669	633	627	592	485	538					
		732	698	695	655	633	618	575	556	532					
		710	698	769	757	698	590	574	521	462	521				
		742	719	742	691	637	625	567	540	512	476				
		740	722	751	692	716	692	657	598	568	544	550			
		758	736	736	702	663	622	602	579	544	575				
		828	769	751	716	728	669	669	568	598	586	615	521		
		779	764	783	752	504	691	652	629	599	556	599	508		
		793	817	811	704	769	710	598	710	609	704	633	598		
		810	802	821	794	736	742	712	691	663	618	666	575	575	
		763	787	917	840	686	692	698	834	633	568	651	503	586	527
		810	816	869	846	729	748	716	691	648	591	655	536	532	448

Average discrepancy (per column)

Average discrepancy (total) = .037

Standard Deviation = .053

Average discrepancy (total) = .037

*Top line indicates actual proportions. **Bottom line indicates theoretical proportions.

TABLE 4

Multiple Category Method
(Decimal points omitted.)

	1	2	3	4	5
Chopin	A* 053	065	112	376	394
	T* 044	094	111	350	421
Cervantes	A 024	118	170	465	223
	T 025	115	172	464	224
Michelangelo	A 024	070	353	506	
	T 023	062	086	305	524
Sarmiento	A 070	170	188	406	165
	T 070	175	188	396	111
Napoleon	A 006	076	100	276	541
	T 021	061	084	306	528
Roosevelt	A 024	041	094	365	476
	T 019	078	072	335	496
San Martín	A 035	141	165	465	194
	T 035	136	181	448	200
Mussolini	A 182	153	224	288	153
	T 169	205	166	304	156
Columbus	A 070	188	194	347	200
	T 078	170	181	381	190
Joan of Arc	A 029	112	170	276	412
	T 036	098	124	368	374
Becquer	A 106	212	123	265	294
	T 125	156	140	308	271
Bolívar	A 047	200	112	424	218
	T 055	151	172	410	212
Hitler	A 182	059	118	294	347
	T 152	126	104	247	371
Beethoven	A 012	076	047	229	635
	T 017	046	066	265	606
Garibaldi	A 100	212	247	365	076
	T 094	240	226	359	081

Average discrepancy (per column)

Average discrepancy (total) = .028

Standard Deviation = .028

Average discrepancy (total) = .022

*A = Actual proportions; T = Theoretical proportions.

The stimuli were presented in four different ways: paired comparisons (15 stimuli), successive intervals (25 stimuli), multiple category method (15 stimuli) and graphic rating scale (25 stimuli). Eight arbitrary interval boundaries were superimposed on the continuum to score the results obtained by using the graphic rating scale method. The 15 names included in the paired comparisons were common to all the other methods. Instructions to the subjects stated clearly that the continuum varied from "extreme lack of interest" to "extreme interest" in knowing the persons indicated by the stimuli.

The experimental population consisted of 170 adults, of both sexes, most of them enrolled in teacher training institutions in Montevideo, Uruguay. The tests were administered to groups of 20 to 30 subjects. To check the accuracy of formula (11) a fictitious example with known σ_i values was prepared. The values obtained by using formula (11) and the real values of σ_i are compared in Table 1.

The following operations were performed: *a*) Frequencies and corresponding proportions were obtained for all the cells, (the actual values for these proportions are given in the upper portion of the cells in Tables 2, 3, 4 and 5). *b*) Cumulative proportions were calculated and the corresponding normal deviates determined. *c*) The L_i and S_i values were computed using formulas (12) and (13), Tables 6 and 7. *d*) The values of σ_i were determined by means of formula (11), Table 6, as follows: *i*) from the X_{ij} values the V_i values were computed, *ii*) the sum of all $(1/V_i)$ values was determined, *iii*) noting that in formula (11) $\sum_i (1/V_i)$ and n are constants, for a stimulus j the value of σ_j was obtained by finding first the value of V_j and then applying formula (11). *e*) Improved estimates of L_i and S_i were obtained by means of formulas (5) and (6), Tables 6 and 7. *f*) The original proportions were reproduced using formula (14). (These values are given in the lower portions of the cells in Tables 2, 3, 4 and 5). The paired comparison data were analyzed using Thurstone's cases III and V, as described by J. P. Guilford (3).

Figure 1 indicates that the relationship between the S_i values obtained by means of the different procedures here described may be interpreted as linear. The slope of the best-fitting line is an indication of the relative dispersion of the S_i values in the two procedures. The paired comparison procedure gives the maximum scatter while the method of successive intervals reduces this scatter to a minimum. This may be due to the actual manner of presenting the stimuli. It should also be remembered that the best determination of the correlations between stimuli. As it will be shown in a future paper there are reasons to believe that some of the stimuli used in this study have substantial correlations among themselves. Consequently, this may explain some of the discrepancies found in this study. Our results are in agreement with those reported by Hevner (4), Saffir (5) and Edwards (1).

TABLE 6

	Values of S_j (Formula 13)				Values of S_j (Formula 6)				Values of σ_j			
	Method of Successive Intervals	Graphic Rating Scale	Multiple Category Method	Paired Comparisons (Case V)	Method of Successive Intervals	Graphic Rating Scale	Multiple Category Method	Paired Comparisons (Case III)	Method of Successive Intervals	Graphic Rating Scale	Multiple Category Method	
Roosevelt	.476	.502	.475	.363	.422	.412	.465	.256	.960	.917	1.000	Roosevelt
Leonardo	.342	.384			.369	.329			1.034	.945		Leonardo
Hitler	-.207	-.112	-.238	-.064	-.038	-.038	-.038	-.065	1.377	1.278	1.532	Hitler
Garibaldi	-.601	-.514	-.584	-.704	-.624	-.522	-.605	-.786	.797	.771	.764	Garibaldi
Marie Antoinette	-.296	-.111			-.298	-.406			1.016	1.074		Marie Antoinette
Mussolini	-.591	-.662	-.592	-.712	-.600	-.680	-.600	-.597	.998	1.109	1.054	Mussolini
Joan of Arc	.242	.106	.179	.218	.223	.080	.119	.214	.988	.966	.976	Joan of Arc
Shakespeare	.521	.500			.611	.457			1.107	.963		Shakespeare
San Martín	-.178	-.058	-.060	-.196	-.297	-.154	-.195	-.175	.778	.806	.782	San Martín
Béquer	-.236	-.255	-.300	-.220	-.234	-.250	-.252	-.265	1.024	1.031	1.179	Béquer
Isabella	-.658	-.760			-.668	-.720			.876	.804		Isabella
Cleopatra	-.349	-.446			-.318	-.440			1.120	1.122		Cleopatra
Dante	-.013	.097			.054	.037			1.114	.911		Dante
Beethoven	.441	.456	.633	.745	.563	.651	.765	1.045	1.116	1.201	1.112	Beethoven
Cervantes	.034	.330	.062	.082	-.151	.188	-.112	.013	.752	.840	.764	Cervantes
Napoleon	.458	.548	.486	.564	.602	.692	.541	.431	1.134	1.137	1.060	Napoleon
Chopin	.162	.214	.186	.325	.279	.342	.245	.393	1.149	1.177	1.085	Chopin
Michelangelo	.371	.411	.464	.604	.446	.484	.535	.629	1.080	1.081	1.074	Michelangelo
Columbus	-.133	-.230	-.280	-.391	-.208	-.273	-.325	-.459	.878	.871	.903	Columbus
Gandhi	-.230	-.217			-.184	-.186			1.121	1.110		Gandhi
Ruben Darío	-.086	-.119			-.111	-.172			.973	.951		Ruben Darío
Madame Curie	.363	.263			.302	.188			.950	.911		Madame Curie
Pasteur	.675	.436			.471	.384			.855	.950		Pasteur
Sarmiento	-.413	-.343	-.284	-.465	-.449	-.332	-.346	-.462	1.011	1.079	.852	Sarmiento
Bolívar	-.094	-.062	-.148	-.148	-.223	-.069	-.224	-.142	.794	.994	.864	Bolívar
Σ	.000	-.002	-.001	.001	-.001	.002	.003	.000	25.002	24.999	15.001	Σ

TABLE 7

Values of I_1 (Formula 12)				Size of Interval (from Formula 12)			
Method	Graphic	Multiple		Method	Graphic	Multiple	
Successive	Rating	Category		Successive	Rating	Category	
Intervals	Scale	Method		Intervals	Scale	Method	
I_1	-2.095	-1.995	-1.637	.332	.296	.706	
I_2	-1.763	-1.659	-.931	.325	.268	.458	
I_3	-1.458	-1.391	-.473	.316	.287	.985	
I_4	-1.122	-1.104	.512	.521	.817		
I_5	-.601	-.287		.555	.516		
I_6	-.046	.229		.472	.467		
I_7	.426	.696		.788	.511		
I_8	1.214	1.207					
Σ	-5.425	-4.264	-2.529				

Value of I_1 (Formula 5)				Size of Interval (Formula 5)			
Method	Graphic	Multiple		Method	Graphic	Multiple	
Successive	Rating	Category		Successive	Rating	Category	
Intervals	Scale	Method		Intervals	Scale	Method	
I_1	-2.076	-1.936	-1.611	.332	.295	.676	
I_2	-1.744	-1.641	-.935	.318	.266	.445	
I_3	-1.426	-1.375	-.490	.304	.276	.956	
I_4	-1.122	-1.099	.466	.507	.800		
I_5	-.615	-.299		.541	.508		
I_6	-.074	.209		.466	.459		
I_7	.392	.668		.783	.509		
I_8	1.175	1.177					
Σ	-5.460	-4.296	-2.570				

TABLE 8
Method of Successive Intervals

	1	2	3	4	5	6	7	8
Roosevelt	(-2.478) (-2.240) (-1.935)	-1.675	-1.216	-1.216	-1.216	-1.559	-.015	.885
Leonardo	(-2.335) (-2.097) (-1.728)	-1.426	-1.024	-1.024	-1.024	-.561	.045	.762
Hitler	(-1.432) (-1.160) (-1.028)	-.803	-.493	-.493	-.493	-.121	.251	1.019
Garibaldi	(-1.801) (-1.563) (-1.160)	-.542	-.148	-.148	-.148	.974	1.347	1.977
Marie Antoinette	-2.097	-1.385	-.904	-.681	-.146	.225	.613	1.316
Mussolini	-1.675	-1.076	-.722	-.459	.073	.490	.863	1.812
Joan of Arc	-2.257	-1.977	-1.799	-1.347	-.927	-.222	.238	.927
Shakespeare	(-2.573) (-2.097) (-1.799)	-1.728	-1.311	-1.311	-1.311	-.739	-.161	.995
San Martín	-2.257	-1.881	-1.426	-1.071	-.407	.240	.803	1.995
Becquer	-2.257	-1.353	-.908	-.703	-.253	.192	.592	1.155
Isabella	-1.739	-1.248	-.842	-.577	-.238	.779	1.426	1.799
Cleopatra	-1.563	-1.316	-.987	-.742	-.240	.235	.662	1.311
Dante	-1.812	-1.563	-1.353	-1.131	-.650	-.210	.251	1.150
Beethoven	(-2.335) (-2.097) (-1.728)	-1.506	-1.122	-1.122	-1.122	-.607	-.174	.616
Cervantes	(-2.424) (-2.186) (-1.881)	-1.311	-.610	-.210	.210	.867	(1.637)	
Napoleon	(-2.335) (-2.097) (-1.881)	-1.426	-1.024	-.616	-.222	.542	-.222	.542
Chopin	-2.097	-1.799	-1.385	-1.216	-.820	-.284	.118	.759
Michelangelo	-2.257	-1.977	-1.799	-1.426	-1.126	-.539	-.028	.762
Columbus	-2.097	-1.799	-1.385	-1.126	-.473	.253	.800	1.468
Gandhi	-1.739	-1.316	-1.103	-.931	-.269	.088	.440	1.248
Ruben Darío	-1.977	-1.799	-1.385	-.974	-.426	.012	.504	1.311
Madame Curie	(-2.335) (-2.097) (-1.977)	-1.563	-.999	-.443	-.043	.1045	.043	1.045
Pasteur	(-2.760) (-2.542) (-2.237)	-1.977	-1.468	-.684	-.164	1.024	-.164	1.024
Sarmiento	-1.616	-1.316	-.999	-.722	-.194	.313	.800	1.616
Bolívar	(-2.335) (-2.097) (-1.616)	-.999	-.284	-.284	-.284	.762	1.616	
n	15	22	23	25	25	25	24	24
Σ	-28.872	-37.110	-31.786	-28.062	-15.023	-1.140	10.661	28.710
Average	-1.925	-1.687	-1.382	-1.122	-.601	-.046	.426	1.196
Dif. Successive Columns	.238	.305	.260	.521	.555	.472	.770	
Σ Total Columns	-52.365	-44.078	-35.958	-28.062	-15.023	-1.140	10.661	30.347
Average (I_1)	-2.095	-1.763	-1.438	-1.122	-.601	-.046	.426	1.214
Size Intervals	.332	.325	.316	.521	.555	.472	.788	

$$\Sigma_{i=1}^n I_i = -5.425$$

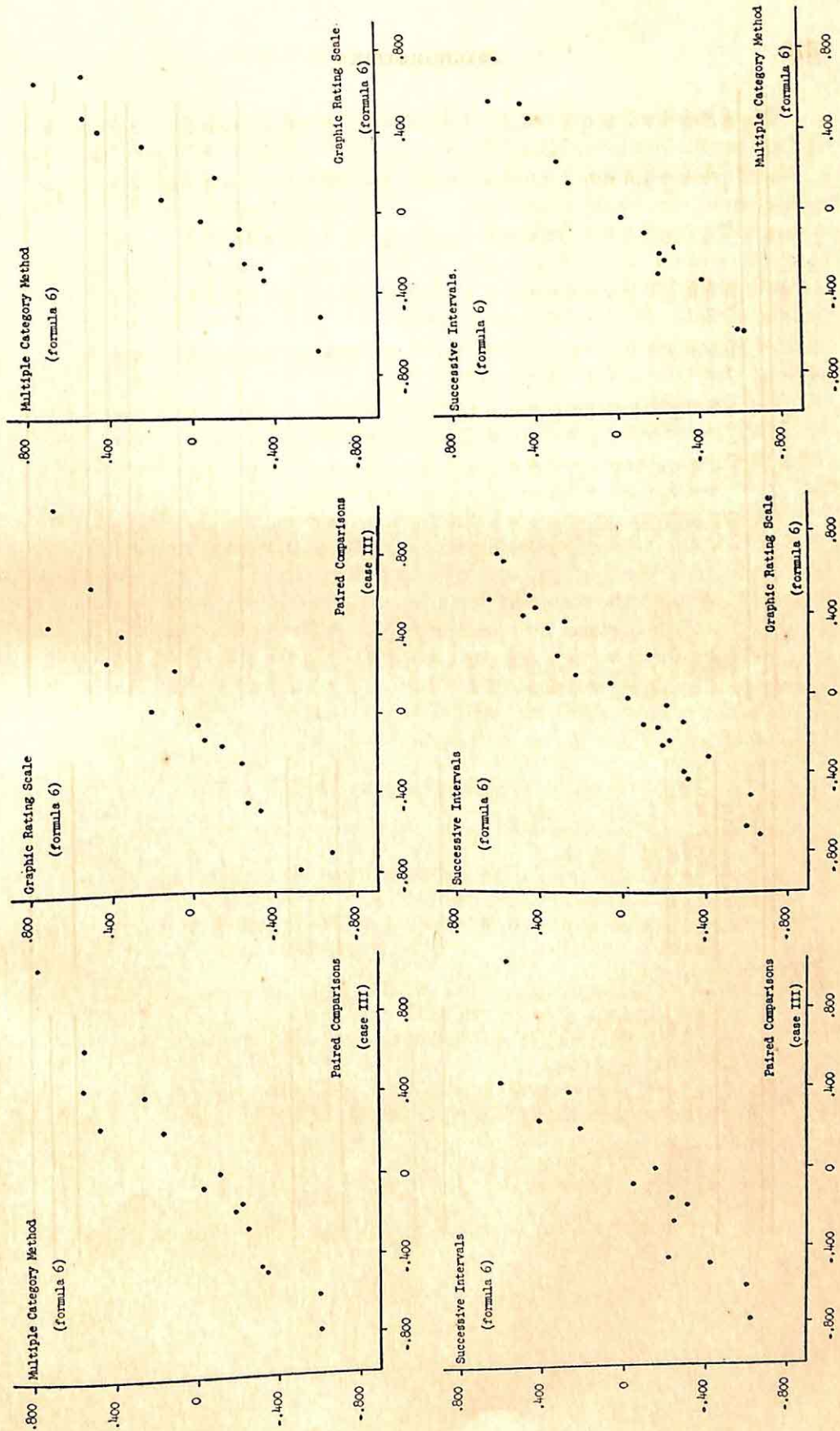


FIGURE 1

The procedure employed to deal with extreme X values is as follows: a) Cells with cumulative proportions above .990 or below .010 were deleted. b) The X values for the remaining cells were calculated and the average for each column was found (Table 8). c) In Table 8 the difference between the averages for successive columns was found; for instance, the difference between columns 3 and 4 was .260. d) This average was used to determine the estimated value corresponding to the deleted cells in column 3; for instance, for the stimulus Pasteur the new value was $-1.977 + (-.260) = -2.237$, and for Roosevelt it was -1.935 . A similar procedure was followed to complete all the missing cells in the table.

The estimated values in Table 8 are given in parentheses (16 values out of 225 for Table 8, 10 out of 225 for graphic rating scale, and 1 out of 75 for multiple category method). As a final check on the operations remember that $\sum S_i = 0$, and $\sum \sigma_i = n$.

The theoretical proportions obtained using formula (14) should agree closely with the actual proportions provided the assumptions used in the development of the method are substantially correct. At the bottom of Tables 2, 3, 4 and 5 the standard deviation and average discrepancies between the actual and theoretical proportions are given. It is clear that the S_i and σ_i values obtained in the manner indicated in this article reproduce the experimental values satisfactorily. Notice that the reproduction of the experimental values using the method here described is better than the one obtained using the paired comparisons method.

REFERENCES

1. Edwards, A. L. The scaling of stimuli by the method of successive intervals. *J. appl. Psychol.*, 1952, **36**, 118-122.
2. Burros, R. H. The estimation of the discriminial dispersion in the method of successive intervals. *Psychometrika*, 1955, **20**, 299-305.
3. Guilford, J. P. *Psychometric methods*. New York: Mc-Graw Hill, 1936.
4. Hevner, H. An empirical study of three psychophysical methods. *J. gen. Psychol.*, 1930, **4**, 191-212.
5. Saffir, M. A. A comparative study of scales constructed by three psychophysical methods. *Psychometrika*, 1937, **2**, 179-198.
6. Thurstone, L. L. A method of scaling psychological and educational tests. *J. educ. Psychol.*, 1925, **16**, 433-451.
7. Thurstone, L. L. Psychophysical analysis. *Amer. J. Psychol.*, 1927, **38**, 368-389.
8. Thurstone, L. L. A law of comparative judgment. *Psychol. Rev.*, 1927, **34**, 273-286.

Manuscript received 1/13/54

Revised manuscript received 1/23/55

A STATISTICAL MODEL FOR RELATIONAL ANALYSIS*

R. DUNCAN LUCE† AND JOSIAH MACY, JR.‡

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

AND

RENATO TAGIURI

HARVARD UNIVERSITY

The diadic relationships existing in a group can be defined in terms of the members' choices, rejections, and their perceptions of being chosen and rejected. The number of possible distinct diads is 45. Formulas are given for computing the expected frequency and variance of the different diadic forms expected, when certain random factors are taken into account. These values must be known if the operation of factors other than the specified random ones is to be studied. Values obtained from two models with different assumptions are compared with empirical values. A simplified treatment is possible for groups with ten or more members.

The student of interpersonal processes often needs to describe and classify in some useful form the relationships between individuals. One such classification is given by relational analysis (2), a method developed in conjunction with a series of studies in interpersonal perception. In this classification the relationship between two persons is described in terms of the *feeling* each has for the other, and the *perception* each has of the other's feeling. More specifically, each member of a well-acquainted group is asked to select those he likes most and those he likes least, and also to guess who likes him most and least. This procedure yields a simple but useful description of the relationship existing between each of the $N(N - 1)/2$ pairs in the group.

Since each subject S_i can choose, reject, or omit any other subject S_j , and can feel chosen, rejected, or omitted by him, nine arrangements are possible of S_i 's feelings and perceptions regarding S_j . We shall define a diad between S_i and S_j as any one of the nine possible arrangements of selections of S_i , combined with any one of the nine possible arrangements of selections of S_j , without regard to order. The number of possible distinct diads is 45.

If S_i 's feeling toward S_j be denoted by 0 for like, 1 for omission and 2 for dislike and, if S_i 's predictions of S_j 's feeling toward him be denoted by

*The present problem emerged from research undertaken as part of a project in interpersonal perception being carried out at the Laboratory of Social Relations at Harvard with the financial aid of the Office of Naval Research (Task Order N5ori—07646).

†Now at Columbia University.

‡Now at Johns Hopkins University.

0 for like, 1 for omission and 2 for dislike, then we can represent some of the 45 possible relationships as follows:

S_i	S_j	S_i	S_j
(11)	(11)	(00)	(00)
(01)	(11)	(00)	(22)
(00)	(01)	(22)	(22)

Legend: the first digit in each bracket corresponds to the feeling, the second to the perception.

Some of the possible diads are well integrated, positive, and realistic; others involve contrary feelings and mistaken perceptions; still others indicate a well-developed negative, and recognized, mutual orientation.

Psychologically important features of a group can be described in terms of the frequency of occurrence of the various diads. It is apparent, however, that given the number of choices, omissions, and rejections, and the number of perceptions of choice, omission, and rejection made by each member of a given group, each diad may be expected to occur a certain number of times by chance alone. To interpret observed data we must know something of these chance distributions, so that we will not attempt to give a psychological interpretation to data which can be explained by the operation of chance alone. When we know which specific diads occur from group to group with greater or less than chance frequency, then we can formulate hypotheses about the possible non-chance factors at work. For these reasons it is important to be able to state the expected frequency of occurrence and the variance of each diad type in a group of given size for an assumed chance model. In previous work (3) estimates of these quantities were obtained by constructing a Monte Carlo robot "group," which was set to match the real group man by man in the number of choices, omissions, and rejections made and in their respective perceptions. This is clearly an unsatisfactory and inefficient method if it can be replaced by a simple mathematical formula.

The purpose of this paper is to present a model in terms of which we can estimate the expected chance frequency and variance of the various diads. It should be borne in mind that the distribution of such frequencies is, probably, often more Poisson than normal.

I. The Model

Several possible "chance" models are conceivable, depending on what we choose to regard as chance. The first one we shall examine corresponds to the case in which the members of a group are regarded as automata, randomly allocating their selections according to fixed probabilities of choosing, rejecting, or omitting every member of the group. Three other

assumptions are made. First, statistical independence is assumed among the different choices, and between the choices and guesses, made by any individual. Second, the choices and guesses of any subject are assumed to be independent of those made by any other subject. Finally, we assume each subject may not choose or guess the same other subject more than once.

For this model, in other words, we assume that the chance occurrence does not include the operation of any psychological factors except those which govern the relative frequencies of the choices and perceptions. In section III we shall discuss a modification of this, in which we assume an S 's perceptions to be conditioned by his choices and rejections.

Let us now proceed with the derivation of the expressions for the expected frequency and variance of each of the diad types. Let S_i 's feeling toward S_j be denoted by 0 for like, 1 for omission, and 2 for dislike. Let S_i 's prediction of S_j 's feeling toward him be denoted by 0 for like, 1 for omission, and 2 for dislike. S_i 's statement of his relationship with S_j will be written $(k_1 k_2)_{ij}$. Then a diad may be denoted $(k_1 k_2)_{ij} (k'_1 k'_2)_{ji}$, where $k_1 = 0, 1$ or 2 etc., and since we do not consider the order, the diads $(k_1 k_2)_{ij} (k'_1 k'_2)_{ji}$ and $(k'_1 k'_2)_{ji} (k_1 k_2)_{ij}$ are identical. We will sometimes distinguish between them for computational reasons, but in general we shall denote either of these diads by $(k_1 k_2) (k'_1 k'_2)$.

For this model we have assumed that each S_i has a fixed probability $P_i(k_1)$ of liking, not mentioning, or disliking each (other) S_j and that this $P_i(k_1)$ is independent of j ; similarly S_i has a fixed probability $Q_i(k'_2)$ of predicting these feelings on the part of each S_j , and this is independent of j and of P_i .

Now let X_{ij} be a random variable which assumes the value 1 if the diad between i and j has some specified value $(k_1 k_2) (k'_1 k'_2)$, and is 0 otherwise. Then

$$X = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n X_{ij} \quad (i \neq j)$$

is a random variable representing the frequency of occurrence of this specified diad in the group. (The following formulas are readily generalized to situations in which any fixed number of categories of questions are answered by the S 's and for which the number of possible responses in each category need only be required to be finite. However, many more than two categories with three or four responses each are not very practical.) Since X_{ij} are all independent,

$$E(X) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n E(X_{ij}) \quad (i \neq j)$$

and

$$E(X_{ij}) = P_i(k_1) Q_i(k'_2) P_j(k'_1) Q_j(k_2),$$

so

$$E(X) = \frac{1}{2} \{ [\sum_i P_i(k_1)Q_i(k_2)] [\sum_i P_i(k'_1)Q_i(k'_2)] - \sum_i P_i(k_1)Q_i(k_2)P_i(k'_1)Q_i(k'_2) \}.$$

And similarly

$$\text{var}(X) = \frac{1}{2} \sum_i \sum_j \text{var}(X_{ij}) \quad (i \neq j)$$

and

$$\text{var}(X_{ij}) = E(X_{ij})[1 - E(X_{ij})],$$

so

$$\begin{aligned} \text{var}(X) = E(X) - \frac{1}{2} \{ [\sum_i P_i^2(k_1)Q_i^2(k_2)] [\sum_i P_i^2(k'_1)Q_i^2(k'_2)] \\ - \sum_i [P_i(k_1)Q_i(k_2)P_i(k'_1)Q_i(k'_2)]^2 \}. \end{aligned}$$

Table 1 shows the observed number of choices, omissions, and rejections, and the number of perceptions of choice, omission, and rejection given by each of the members of a ten-man group. The data used as an example were

TABLE 1
Observed Frequencies of Different Feelings
and Perceptions in a Ten-Man Group

Sub- ject	Feeling			Perception		
	k = 0	k = 1	k = 2	k' = 0	k' = 1	k' = 2
1	3	5	1	3	5	1
2	3	2	4	5	1	3
3	6	0	3	4	3	2
4	2	6	1	2	5	2
5	2	4	3	3	3	3
6	4	3	2	1	5	3
7	4	4	1	2	6	1
8	3	4	2	3	4	2
9	2	5	2	2	5	2
10	3	5	1	1	7	1

TABLE 3

Conditional Probabilities $Q(k_2 k_1)$ for the Ten-Man Group			
k_1	$Q(0 k_1)$	$Q(1 k_1)$	$Q(2 k_1)$
0	0.63	0.35	0.02
1	0.24	0.67	0.10
2	0.07	0.40	0.53

TABLE 2

Observed and Expected Frequencies of Congruous
and Non-Congruous Diads; Model with Perception
Not Contingent upon Feeling

Diad Type	Observed	Expected
Bilateral Congruency	20	6.87
Unilateral Congruency	22	21.37
No Congruency	3	16.76
Chi square = 36.32 d.f. = 2 p < 0.001		

TABLE 4

Observed and Expected Frequencies of Congruous
and Non-Congruous Diads; Model with Perception
Contingent upon Feeling

Diad Type	Observed	Expected
Bilateral Congruency	20	17.63
Unilateral Congruency	22	21.56
No Congruency	3	6.07
Chi square = 1.88 d.f. = 2 p = 0.40		

obtained at the end of the last meeting of a series of twelve sessions conducted by a psychoanalyst. The nature of the meetings was a modified form

of group therapy where the members met to "discuss principles of group psychology particularly as these relate to self understanding." The procedure consisted of asking the members to indicate those others in the group they "liked most" and "least" as well as to guess who would name them as liking them "most" and "least."

Analysis of the data in terms of the particular composition of each of the $N(N - 1)/2$ diads gives the observed frequency of each diad type and, in terms of these, describes the group. For example, the diad (00)(00) has an observed frequency of six, while (02)(20) does not occur; the figures below show that these frequencies are different from the expected value predicted by the chance model.

<i>Diad</i>	<i>Observed</i>	<i>Expected</i>	<i>Variance</i>
(00)(00)	6	0.48	0.46
(02)(20)	0	0.50	0.50

The first diad, in which both subjects like each other and predict being liked by the other, occurs more often than expected by chance; the other, in which feelings are not mutual but are accurately predicted, occurs less often than expected, but not significantly so.

In general we have found that there is a significant discrepancy between observed frequencies and those predicted by this chance model. This indicates that there are factors operating other than those we have assumed in this model, and the differences do suggest the nature of some of these factors.

Let us further exemplify the use of the model. It is quite apparent that, in general, the feeling we hold for a person is *congruent* with the feeling we *perceive* that person holds for us. This tendency is quite apparent in all of our data. Thus member S_i tends to choose and feel chosen by S_i , or to dislike and feel disliked by S_i , etc. Is this tendency sufficiently consistent that diads containing such congruencies between feelings and perceptions would exceed chance, while others would fall below chance? The present model permitted us to test such hypotheses by supplying us with an acceptable chance baseline. The data for the ten-man group mentioned above will be used to illustrate this point. We will separate the diads into three groups. In the first we will put all diads in which feeling and perception are the same for both members: (00)(00), (00)(11), (00)(22), etc. In the second, we will put those diads in which this is true for only one member: (00)(01), (00)(12), etc.; in the third we will put all diads in which this holds for neither member: (01)(12), (10)(21), etc. If our conjecture is right, the first class should contain more cases than expected, and the third class fewer than expected, on the basis of chance alone. The figures presented in Table 2 show that the differences are as predicted; the probability of this occurring with the chance model is less than 0.001.

II. Model with Perceptions Dependent upon the Subject's Feelings

We may now modify the basic chance model by incorporating various hypotheses about the group, to see whether these additional factors will explain the observed results.

We have said above that, in all groups, we have observed a strong tendency for a member to predict that others feel toward him whatever he feels toward them, and we have exhibited this tendency in one group. It seems reasonable to ask whether this tendency alone accounts for the deviation from chance. We shall, therefore, investigate how well a model with this modification accounts for the data.

We shall assume again that each S responds in an independent manner, with probability $P_i(k)$ of liking, omitting, or disliking any other subject. However, we shall now assume that this choice conditions his prediction of another's feeling toward him, so that his probability of predicting a given response on the part of another member is not $Q_i(k')$, as before, but is $Q_i(k' | k)$. The expressions for the expected value for the occurrence of any diad and the variance take similar forms to those presented above, with this conditional probability used for Q_i . In the case of the group used here as an example, we do not have sufficient data to estimate the conditional probabilities individually for each member, so we shall use one set of such probabilities $Q(k' | k)$ for all the members, estimating the values from the data for all members combined. This simplification is not necessary in general but will be used for the example. With this assumption, the expected frequency of occurrence of a given diad reduces to

$$E(X) = [\frac{1}{2}Q(k_2 | k_1)Q(k'_2 | k'_1)] [\sum_i P_i(k_1) \sum_i P_i(k'_1) - \sum_i P_i(k_1)P_i(k'_1)]$$

and the expression for the variance is similarly simplified.

For our group, the conditional probabilities observed are shown in Table 3. If we now combine the diads as we did in Table 2, and compare the frequencies observed and the frequencies predicted using the conditional probabilities, we can observe a striking improvement in agreement. (Cf. Table 4 and compare with Table 2).

Using the chi-square test, we see that there is a probability of about 0.40 that the value of chi-square observed would be exceeded if the hypothesis were true. We can on this evidence neither accept nor reject the hypothesis that the observed frequencies of these diad types are accounted for by a chance model with predictions conditioned by feelings; but the improvement in fit is striking and suggests that a large part of the observed distribution of diad types is due to such contingency. This example illustrates the use of such baseline models in the study of the meaning of the observed frequency distribution for the diad types.

It is obvious that other hypotheses about the group could be tested by

constructing a similar model and examining the observed frequencies to determine how much of the variation is accounted for by such a model. The principle in all cases is the same; a model is constructed which assumes that the members of the group are automata acting at random, with probabilities governed by the particular hypotheses at hand. The expected frequencies obtained from this model are then used to investigate the group and to determine whether we have reason to believe that other psychological processes are at work beyond those assumed in the model. These hypotheses must be chosen with care, however, in order to yield a model which is mathematically tractable and which leads to a practical amount of computational labor.

III. Simplifications for Large Groups

In the models developed above, we have allowed the probabilities P_i and Q_i to be different for each member of the group. This leads to lengthy calculations for large groups. For groups larger than 10, however, we may introduce a simplification which greatly reduces this labor by using the mean value over all members of the group for the value of P_i ; thus each member is described by the same probabilities; thus summations are no longer necessary. In the case of the first model mentioned above, if the mean value of $P_i(k_1)$ is denoted by $P(k_1)$, and the mean of $Q_i(k_2)$ by $Q(k_2)$, we may then write the expected value $E(X)$ as

$$E'(X) = [n(n-1)/2][P(k_1)Q(k_2)P(k'_1)Q(k'_2)],$$

and the expression for the variance is similarly simplified.

Let us examine the error involved in this approximation. Let

$$A(X) = \sum_i P_i(k_1)Q_i(k_2) - nP(k_1)Q(k_2)$$

and

$$A'(X) = \sum_i P_i(k'_1)Q_i(k'_2) - nP(k'_1)Q(k'_2)$$

and

$$B(X) = \sum_i P_i(k_1)Q_i(k_2)P_i(k'_1)Q_i(k'_2) - nP(k_1)Q(k_2)P(k'_1)Q(k'_2).$$

Then it can easily be shown that if $E(X)$ is the expected value previously calculated using the individual probabilities, and $E'(X)$ is the expected value given above,

$$\begin{aligned} E(X) - E'(X) &= \frac{1}{2}[A(X)A'(X) + nA(X)P(k'_1)Q(k'_2) \\ &\quad + nA'(X)P(k_1)Q(k_2) - B(X)]; \end{aligned}$$

and so if we use $D(X) = [E(X) - E'(X)]/E'(X)$ as a measure of the error,

$$\begin{aligned} D(X) &= \frac{A(X)A'(X)}{n(n-1)P(k_1)Q(k_2)P(k'_1)Q(k'_2)} + \frac{A(X)}{(n-1)P(k_1)Q(k_2)} \\ &\quad + \frac{A'(X)}{(n-1)P(k'_1)Q(k'_2)} - \frac{B(X)}{n(n-1)P(k_1)Q(k_2)P(k'_1)Q(k'_2)}. \end{aligned}$$

Now it has been found from experience that $A(X)/[P(k_1)Q(k_2)]$ and $A'(X)/[P(k'_1)Q(k'_2)]$ are less than 2, and in almost all cases very near 1 for the groups encountered in practice. $B(X)/[(n-1)P(k_1)Q(k_2)P(k'_1)Q(k'_2)]$ is less than 1, and in almost all cases less than $1/2$; so in practice this error $D(X)$ is less than $5/n$ for n greater than 5. In almost all cases this turns out to be a very liberal estimate of the error; for example, in the group of 10 used earlier, typical errors are

$$D[(00)(00)] = 0.00738 = .74\%$$

$$D[(00)(01)] = 2.28\%$$

$$D[(00)(02)] = 0.74\%$$

For the model with $Q_i(k_2)$ given by $Q(k_2 | k_1)$ for all members, the error in replacing the P_i by P is even less. In this case, A and $A'(X)$ are 0, and the error is then less than $1/n$.

This simplification is particularly useful because it introduces the least error for large groups, where it is most needed to simplify the calculation.

IV. Summary

Relational analysis defines the diadic relationship existing between pairs of members of a group in terms of their choices, their rejections, and their perceptions of being chosen and rejected. The number of possible diads is 45. In order to interpret the results of an experiment, we must have knowledge of the expected occurrence of the various diads on a chance basis, when only certain specified processes govern the chance distribution of diads.

This paper discusses the construction of models which give the expected value and variance of the diads, when certain assumptions are made as to the random factors operating in the group. In general, the assumptions are that only very simple psychological factors are operating in the group, and that the occurrence of the various diad forms is the result of chance operating within the restriction of these factors. The observed data are then examined to determine whether the chance model accounts for the distribution of diad types, or whether additional psychological processes must be postulated. Models such as these are essential for testing various hypotheses about interaction in the group since they provide a method for setting up and testing a null hypothesis by the usual statistical methods.

The models discussed in the paper were constructed on the assumption that choices and predictions were independent from member to member, and under the assumption that prediction was conditioned by choice in any pair as well as the assumption that choice and prediction were independent. The first of these assumptions was shown to account for a large part of the observed variation in diad frequency. Simplified assumptions which are valid for large groups were also discussed.

Models such as the second one discussed in this paper are typical of a large variety of models which could be constructed to test various hypotheses about the sources of the variation of frequency of the diad types.

REFERENCES

1. Feller, W. An introduction to probability theory and its application. Vol. 1. New York: Wiley, 1950.
2. Tagiuri, R. Relational analysis: An extension of sociometric method with emphasis upon social perception. *Sociometry*, 1952, 15, 91-104.
3. Tagiuri, R., Blake, R. R., and Bruner, J. S. Some determinants of the perception of positive and negative feelings in others. *J. abnorm. soc. Psychol.*, 1953, 48, 585-592.

Manuscript received 7/7/53

Revised manuscript received 11/22/54

PSYCHOMETRIC SOCIETY

Statement of Receipts and Disbursements for Fiscal Year
Ended June 30, 1955

RECEIPTS (Dues)		
Year	Members	Student Members
1957	1 (\$7 each)	
1956	3 (\$7 each)	
1955	132 (\$7 each)	146 (\$4 each)
1954	66 (\$7 each)	19 (\$4 each)
1953	5 (\$7 each)	
	<u>507</u>	<u>65</u>
		\$3809.00

Received with Dues for Corporation Publications 7.20
Mailing List Rental 25.00
Proceeds of 1954 Joint Dinner with APA Division 5 23.13
Miscellaneous15

Total Receipts \$3864.18

DISBURSEMENTS

Psychometric Corporation (90% of dues) \$3128.10

Miscellaneous Disbursements:
Corporation for Publications . . . \$ 7.20
Mimeographing and Printing . . . 67.77
Postage 70.92
Secretarial Services 165.02
Addressing and Mailing 34.79
Bank Charges 6.79

Total Disbursements 352.19
\$3780.59

BALANCE

Reported Bank Balance, July 1, 1954 \$1011.68
Actual Bank Balance, July 1, 1954 \$1003.91
Receipts 1864.18
Disbursements, 1954-1955 \$1088.39
Bank Balance, June 30, 1955 3780.59
\$1067.80

PSYCHOMETRIC CORPORATION

Statement of Receipts and Disbursements for Fiscal Year
Ended June 30, 1955

Reported Bank Balance, July 1, 1954	\$6322.53
Actual Bank Balance, July 1, 1954	\$6320.72

RECEIPTS

Subscriptions (less agency discounts) . . . \$5439.50
Psychometric Society (90% of dues) . . . 3428.10
Sale of Back Issues (less discounts) . . . 1104.23
Sale of Monographs 5-8 (less discounts) . . 546.00
Royalties from U. Chi. Press (1953-1954) . . 1.38
Reprints, Etchings, and Alterations . . . 206.11
For Monographs 2-4 (less discounts) . . . 117.95
Overpayments 12.35

Total Receipts \$11155.92

DISBURSEMENTS

Printing Psychometrika,
Volume 19, No. 2, through 20, No. 1 . . \$5030.58
Reprints, Etchings, and Alterations . . . 197.86
Stipend of Assistant Editor,
Volume 18, No. 4, through 19, No. 2 . . . 246.00
Secretarial Services 456.81
Stationery and Postage 271.15
Loan Repayments in Full 1100.00
For Monographs 2-4 105.75
Miscellaneous 75.60

Total Disbursements \$7483.75

BANK BALANCE, June 30, 1955 \$9992.89

ESTIMATED OBLIGATIONS

Printing Psychometrika,
Volume 20, Nos. 2, 3, 4 \$3900.00
Stipend of Assistant Editor,
Volume 19, No. 3, through 20, No. 4 . . . 544.00
Secretarial Services 750.00
Handling Charges on Back Issues and
Monographs through 6/30/55 825.00

Total Obligations \$6019.00

BANK BALANCE (less obligations), June 30, 1955 \$3973.89

INDEX FOR VOLUME 20

AUTHOR

- Binder, Arnold, "The Choice of an Error Term in Analysis of Variance Designs." 29-50.
- Boguslavsky, G. W., "A Mathematical Model for Conditioning." 125-138.
- Brogden, Hubert E., "Least Squares Estimates and Optimal Classification." 249-252.
- Burros, Raymond H., "The Estimation of the Discriminal Dispersion in the Method of Successive Intervals." 299-305.
- Denton, J. C. (With Calvin W. Taylor), "A Factor Analysis of Mental Abilities and Personality Traits." 75-81.
- Durost, Walter M., "ANNE ANASTASI, *Psychological Testing*." A Review. 261-262.
- Eysenck, H. J., "ETS, *Kit of Selected Tests for Reference Aptitude and Achievement Factors*." A Review. 168-169.
- Frankel, Stanley, "On the Design of Automata and the Interpretation of Cerebral Behavior." 149-162.
- Gibson, W. A., "An Extension of Anderson's Solution for the Latent Structure Equations." 69-73.
- Gourlay, Neil, "*F-Test Bias for Experimental Designs in Educational Research*." 227-248.
- Gourlay, Neil, "*F-Test Bias for Experimental Designs of the Latin Square Type*." 273-287.
- Green, Bert F., "J. P. GUILFORD, *Psychometric Methods (2nd Ed.)*." A Review. 163-165.
- Guilford, J. P., "Louis Leon Thurstone." Obituary. 263-265.
- Guttman, Louis, "A Generalized Simplex for Factor Analysis." 173-192.
- Guttman, Louis, "Reliability Formulas for Noncompleted or Speeded Tests." 113-124.
- Harris, Chester W., "Characteristics of Two Measures of Profile Similarity." 289-297.
- Harris, Chester W., "Separation of Data as a Principle in Factor Analysis." 23-28.
- Hormaeche, M. (With H. J. A. Rimoldi), "The Law of Comparative Judgment in the Successive Intervals and Graphic Rating Scale Methods." 307-318.
- Jenkins, W. L., "An Improved Method for Tetrachoric r ." 253-258.
- Lord, Frederic M., "Equating Test Scores—A Maximum Likelihood Solution." 193-200.
- Lord, Frederic M., "Sampling Fluctuations Resulting from the Sampling of Test Items." 1-22.
- Lorge, Irving (With Herbert Solomon), "Two Models of Group Behavior in the Solution of Eureka-Type Problems." 139-148.
- Luce, R. Duncan (With Josiah Macy, Jr. and Renato Tagiuri), "A Statistical Model for Relational Analysis." 319-327.
- Macy, Josiah, Jr. (With R. Duncan Luce and Renato Tagiuri), "A Statistical Model for Relational Analysis." 319-327.
- Michael, William B., "RAYMOND B. CATTELL, *Factor Analysis: An Introduction and Manual for the Psychologist and Social Scientist*." A Review. 166-168.
- Payne, M. Carr, Jr. (With Leonard Staugas), "An IBM Method for Computing Intraserial Correlations." 87-92.
- Psychometric Corporation, Report of the Treasurer. 329.
- Psychometric Society, Report of the Treasurer. 329.
- Rao, C. Radhakrishna, "Estimation and Tests of Significance in Factor Analysis." 93-111.
- Restle, Frank, "Axioms of a Theory of Discrimination Learning." 201-208.
- Rimoldi, H. J. A. (With M. Hormaeche), "The Law of Comparative Judgment in the Successive Intervals and Graphic Rating Scale Methods." 307-318.
- Solomon, Herbert (With Irving Lorge), "Two Models of Group Behavior in the Solution of Eureka-Type Problems." 139-148.

- Staugas, Leonard (With M. Carr Payne, Jr.), "An IBM Method for Computing Intraserial Correlations." 87-92.
- Tagiuri, Renato (With R. Duncan Luce and Josiah Macy, Jr.), "A Statistical Model for Relational Analysis." 319-327.
- Taylor, Calvin W. (With J. C. Denton), "A Factor Analysis of Mental Abilities and Personality Traits." 75-81.
- Thurstone, L. L., "Sir Godfrey Thomson." Obituary. 171-172.
- Tucker Ledyard R, "The Objective Definition of Simple Structure in Linear Factor Analysis." 209-225.
- Tucker, Ledyard R, "Psychometric Theory: General and Specific." 267-271.
- Tucker, Ledyard R, "A Rational Curve Relating Length of Rest Period and Length of Subsequent Work Period for an Ergographic Experiment." 51-61.
- Webster, Harold, "C. RADHAKRISHNA RAO, *Advanced Statistical Methods in Biometric Research.*" A Review. 165-166.
- Welsh, George Schlager, "A Tabular Method of Obtaining Tetrachoric r with Median-Cut Variables." 83-85.
- Winer, Ben J., "A Measure of Interrelationship for Overlapping Groups." 63-68.
- Wrigley, Charles, "BENJAMIN FRUCHTER, *Introduction to Factor Analysis.*" A Review. 259-260.



